

# Compressive Sensing Based Joint Activity and Data Detection for Grant-Free Massive IoT Access

Yikun Mei, Zhen Gao, Yongpeng Wu, Wei Chen, Jun Zhang, Derrick Wing Kwan Ng, *Fellow, IEEE*, Marco Di Renzo, *Fellow, IEEE*

**Abstract**—Massive machine-type communications (mMTC) are poised to provide ubiquitous connectivity for billions of Internet-of-Things (IoT) devices. However, the required low-latency massive access necessitates a paradigm shift in the design of random access schemes, which invokes a need of efficient joint activity and data detection (JADD) algorithms. By exploiting the feature of sporadic traffic in massive access, a beacon-aided slotted grant-free massive access solution is proposed. Specifically, we spread the uplink access signals in multiple subcarriers with pre-equalization processing and formulate the JADD as a multiple measurement vectors (MMV) compressive sensing problem. Moreover, to leverage the structured sparsity of uplink massive access signals among multiple time slots, we develop two computationally efficient detection algorithms, which are termed as orthogonal approximate message passing (OAMP)-MMV algorithm with simplified structure learning (SSL) and accurate structure learning (ASL). To achieve accurate detection, the expectation maximization algorithm is exploited for learning the sparsity ratio and the noise variance. To further improve the detection performance, channel coding is applied and successive interference cancellation (SIC)-based OAMP-MMV-SSL and OAMP-MMV-ASL algorithms are developed, where the likelihood ratio obtained in the soft-decision can be exploited for refining the activity identification. Finally, the state evolution of the proposed OAMP-MMV-SSL and OAMP-MMV-ASL algorithms is derived to predict the performance theoretically. Simulation results verify that the proposed solutions outperform various state-of-the-art baseline schemes, enabling low-latency random access and high-reliable massive IoT connectivity with overloading.

The work of Z. Gao was supported in part by the Natural Science Foundation of China (NSFC) under Grant 62071044; in part by the Beijing Natural Science Foundation under Grant L182024; and in part by the Basic Science Center for Autonomous Intelligent Unmanned Systems under Grant 62088101, and in part by the Science and Technology Innovation Plan from Beijing Institute of Technology. The work of Y. Wu was supported in part by the National Key R&D Program of China under Grant 2018YFB1801102 and NSFC under Grant 62071289. The work of D. W. K. Ng was supported by funding from the UNSW Digital Grid Futures Institute, UNSW, Sydney, under a cross-disciplinary fund scheme and by the Australian Research Council's Discovery Project under Grant DP210102169. (*Corresponding author: Zhen Gao.*)

Y. Mei and Z. Gao are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mails: meiyikun@bit.edu.cn; gaozhen16@bit.edu.cn).

Y. Wu is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yongpeng.wu@sjtu.edu.cn).

W. Chen is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China (e-mail: weich@bjtu.edu.cn).

D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2025, Australia (e-mail: w.k.ng@unsw.edu.au).

M. D. Renzo is with the Laboratoire des Signaux et Systèmes, CNRS, CentraleSupélec, Univ Paris Sud, Université Paris-Saclay, 91192 Paris, France (e-mail: marco.direnzo@centralesupelec.fr).

**Index Terms**—Compressive sensing, grant-free massive access, orthogonal approximate message passing, multiple measurement vectors, successive interference cancellation.

## I. INTRODUCTION

Next-generation wireless communications promise the unprecedented connectivity for myriad Internet-of-Things (IoT) devices, which compose the prime driver of massive machine-type communications (mMTC) [1], [2]. In stark contrast to the conventional downlink-dominated human-type communications (HTC), the emerging mMTC are uplink-dominated having the features of sporadic traffic with low rates and short packets [1]–[4]. With this premise, mMTC compel a paradigm shift in the design of uplink random access procedures, which previously focused on HTC while were rarely prioritized for massive machine-type IoT devices [4].

The current random access protocols, including long term evolution-advanced (LTE-A) and the fifth-generation new radio (5G NR) release 16, adopt the conventional grant-based access procedure [5]. In particular, grant-based random access usually requires a complicated handshaking protocol, through which a base station (BS) allocates orthogonal radio resources to different users for uplink transmission. However, such grant-based protocols suffer from low access efficiency in mMTC, since the payload size can be much smaller than the amount of the signaling overhead. Moreover, since the IoT devices can be massive, the contention for resources request will lead to inevitable collisions, which can dramatically increase the access latency [6], [7]. Therefore, an efficient massive random access paradigm is needed for enabling low-latency and high-reliability mMTC.

### A. Related Work

To accommodate the challenging requirements of mMTC, the grant-free massive access paradigm has been emerging as a promising solution and has drawn great attention [8], [9]. In grant-free random access schemes, users directly transmit their data to the BS without performing a sophisticated handshaking process. In this way, complicated control signaling procedures can be avoided, the access latency can be substantially reduced, and more radio resources can be saved for payload transmission. On the other hand, since the number of potential IoT devices in mMTC is generally massive, it is impossible to allocate orthogonal channel resources to all devices for facilitating data detection at the BS [10], [11]. Fortunately, IoT devices, in most instances, remain inactive for energy saving and other reasons (e.g., data collection and infrequent alarm).

In other words, only a small number of devices are active and need to access the BS at any given time.

In view of this fact, most grant-free access protocols adopt limited non-orthogonal radio resources to serve massively deployed IoT devices [14]–[21], where the detection of devices' activity and data is a challenging problem to be solved. Due to the feature of sporadic traffic in mMTC, the joint activity and data detection (JADD) procedure can be formulated as a sparse signal recovery problem, which can be handled by using compressive sensing (CS) algorithms. In fact, CS-based multi-user detectors (MUD) can date back to code division multiple access (CDMA) systems. In [12], for instance, MUDs based on the sparsity-exploiting maximum a posteriori probability (MAP) criterion were proposed for sparse CDMA systems. Besides, considering that the sparsity can be time-varying, CS-based nonlinear detectors and conventional linear detectors were combined for improving robustness [13]. Also, with the advent of mMTC, grant-free random access in conjunction with CS-based detectors were conceived for JADD in [14]–[21]. In general, these CS-based detection algorithms can be divided into two categories: greedy-based algorithms and Bayesian inference-based algorithms [22]–[26]. The former category was developed from the CS greedy algorithms including the orthogonal match pursuit (OMP) algorithm [22], the subspace pursuit (SP) algorithm [23], etc., which heuristically find the most correlated atoms in each iteration. Furthermore, to exploit the temporal partial correlation of device activity, a modified OMP algorithm and a modified SP algorithm were proposed in [14] and [15], respectively, where the estimated support was exploited as prior knowledge for the next detection. Moreover, considering that the temporal correlation of device activity typically remains unchanged during one frame, the structured iterative support detection algorithm and modified adaptive SP algorithm were proposed in [16] and [17], respectively, whereby the block sparsity of random access signals in multiple slots was leveraged for performance improvement. However, [14]–[17] fail to utilize the a priori distribution of the transmitted constellation symbols, and the calculation of matrix inversions that are needed in those algorithms pose an exceedingly high computational complexity when the number of devices becomes large.

Another category of CS-based random access detection algorithms are built on the framework of Bayesian inference, which includes the approximate messaging passing (AMP) algorithm [24], the sparse Bayesian learning (SBL) algorithm [25], the expectation propagation (EP) algorithm [26], and etc. A common feature of them lies in the fact that the adopted estimators are designed based on the posterior distribution, which often depends on the a priori distribution of devices' symbols. Due to their flexibility and accuracy, they have become popular recently. In [19], for example, the modified SBL algorithms were applied for JADD, by making use of the sparse device activity and the prior knowledge of the signals' constellation. By resorting to the expectation-maximization (EM) algorithm, the knowledge of device activity is not required by the proposed algorithms. However, the associated computational complexity can be high since the SBL algorithm requires the computation of a matrix inversion in each

iteration. By subtly using the likelihood ratio as the judgment metric for JADD, a MAP-based iterative detection scheme was proposed in [20]. However, the noise variance and probability of devices' activity were assumed to be perfectly known at the BS, which can be impractical. To effectively reduce the computational complexity and speed up the convergence, an AMP-based detection scheme was proposed in [21], where the hyper-parameters of device activity can be learned through the EM algorithm. In [21], in addition, the sensing matrix consists of channel coefficients and spreading sequences. This indicates that the channel state information (CSI) is assumed to be available in advance at the BS. Moreover, the adopted spreading sequences require the elements to be independent and identically distributed (i.i.d), which results in the design difficulty. To address those issues, our works aim at designing a practical JADD that guarantees high efficiency and accuracy.

### B. Our Contribution

In this paper, we propose a beacon-aided slotted grant-free massive access scheme for mMTC and investigate the JADD for application to orthogonal frequency division multiplexing (OFDM) systems<sup>1</sup>. Specifically, due to the sporadic traffic of machine-type IoT devices, the JADD problem at the BS can be formulated as a CS problem and can be solved by using the orthogonal AMP (OAMP) algorithm [27]. Moreover, the structured sparsity of uplink access signals in several successive OFDM symbols can be exploited by leveraging the multiple measurement vectors (MMV) CS algorithm for performance enhancement. However, the conventional OAMP algorithm is limited to the single measurement vector (S-MV) CS problem and is incapable of solving the MMV CS problem with structured sparsity. To this end, we propose two detection algorithms that are referred to as OAMP-MMV algorithms with simplified structure learning (SSL) and accurate structure learning (ASL). In both of them, the prior knowledge of discrete constellation symbols is utilized, and EM algorithms are employed to estimate the sparsity ratio and noise variance for reliable detection performance. Even for the overloading case, the proposed OAMP-MMV algorithms can still achieve effective JADD. Meanwhile, the state evolution (SE) is derived for theoretically analyzing the performance of our proposed algorithms. Furthermore, channel coding and successive interference cancellation (SIC) are integrated into the proposed OAMP-MMV-SSL and OAMP-MMV-ASL algorithms for performance enhancement, where the likelihood ratio (LLR) is utilized for improving the activity identification. The main contributions made by this paper can be summarized as follows.

- **Beacon-aided slotted grant-free massive access scheme:**

Previous works, e.g., [14]–[21], mainly considered that the channels of all devices are known at the BS. In contrast, we remove this assumption and propose a beacon-aided random access scheme. After listening to the beacon that is periodically broadcast by the BS, the IoT devices perform synchronization, power control, and channel estimation. Then,

<sup>1</sup>This article was presented in part at the International Conference on UK-China Emerging Technologies, Glasgow, UK, August 2020.

a pre-equalization processing is introduced at the devices so that the uplink channels required by the BS are effectively avoided.

- **OAMP-MMV-SSL algorithm:** Exploiting the structured sparsity in several successive time slots, we develop an OAMP-MMV-SSL algorithm, where the SSL strategy is proposed to improve the performance of activity detection. In particular, the discrete a priori distribution is assumed for characterizing the transmitted constellation symbols, and EM algorithm is incorporated to learn the unknown sparsity ratio and the noise variance. Besides, the SE is derived for characterizing the algorithm's performance.

- **OAMP-MMV-ASL algorithm:** To fully exploit the aforementioned structured sparsity, we further propose an OAMP-MMV-ASL algorithm, where an ASL strategy with more sophisticated a priori distribution is considered. Particularly, by establishing a graphical model, the structured sparsity can be accurately described in the a priori distribution. We show that the OAMP-MMV-ASL algorithm can achieve better performance than OAMP-MMV-SSL algorithm at the cost of increasing the computational complexity.

- **SIC-based JADD:** In contrast to most prior works, e.g., [14], [16]–[21], which assume uncoded random access, our proposed access scheme takes channel coding into account and develop SIC-based OAMP-MMV-SSL and OAMP-MMV-ASL algorithms. Particularly, the LLR obtained in soft-decision is exploited for refining the active device identification.

Compared to the previous study [21], our proposed schemes are developed from the OAMP algorithm rather than the AMP algorithm, where the OAMP algorithm is more robust than the traditional AMP algorithm and has the less requirements on the sensing matrix. A similar point lies in that both [21] and our work employ the EM method to update the sparsity ratio. However, our proposed OAMP-MMV-ASL algorithm considers the more sophisticated a priori distributions than the method proposed in [21] and resorts to sum-product rule for learning the sparsity structure, which has been verified by the numerical results to have the more accurate JADD performance. Besides, the noise variance is assumed to be known in [21], while we remove this impractical constraint and apply the EM algorithm to learn the unknown noise variance. Last but not least, the integration with channel coding and SIC is not considered in [21], and we will show that the proposed SIC-based OAMP-MMV algorithms effectively improve the performance compared with the OAMP-MMV algorithms without SIC processing.

The rest of the paper is organized as follows. Section II describes the system model of grant-free random access. Section III introduces our proposed scheme and algorithms for JADD, and how SIC and channel coding are integrated into OAMP-MMV algorithms. In Section IV, we analyze the performance of the proposed algorithms by focusing on the SE and computational complexity. Simulation results are given in Section V to demonstrate the superiority of our proposed algorithm. Finally, conclusions are drawn in Section VI.

*Notation:* The boldface lower and upper-case symbols denote column vectors and matrices, respectively.  $(\cdot)^*$ ,  $(\cdot)^T$ ,  $(\cdot)^H$ ,  $(\cdot)^{-1}$ ,  $E[\cdot]$  and  $\text{var}[\cdot]$  denote the conjugate, transpose, conju-

gate transpose, inversion, expectation and variance operators, respectively.  $\|\mathbf{a}\|_p$  is the  $\ell_p$  norm of vector  $\mathbf{a}$ .  $\text{diag}(\mathbf{a})$  is a diagonal matrix with the elements of vector  $\mathbf{a}$  on its diagonal.  $\mathbf{I}_N$  denotes the identity matrix of size  $N \times N$ , and  $\mathbf{O}_{M \times N}$  denotes the null matrix of size  $M \times N$ .  $\mathbf{0}_N$  denotes the vector of size  $N$  with all the elements being zeros, and  $\emptyset$  denotes the empty set.  $\mathbf{A}_{m,:}$  and  $\mathbf{A}_{:,n}$  are the  $m$ -th row vector and the  $n$ -th column vector of matrix  $\mathbf{A} \in \mathbb{C}^{M \times N}$ , respectively.  $\mathcal{CN}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Gamma})$  denotes the complex Gaussian distribution of a random vector  $\mathbf{x}$  with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Gamma}$ .  $\Re\{\cdot\}$  denotes the real part of the corresponding arguments, and  $\text{card}\{\cdot\}$  denotes the cardinality of a set.

## II. SYSTEM MODEL

We consider a typical uplink grant-free massive access scenario with mMTC, where OFDM is adopted to overcome the time dispersive channels. As shown in Fig. 1(a), a single-antenna BS serves  $K$  single-antenna IoT devices, where  $K$  can be very large but only  $K_a$  ( $K_a \ll K$ ) devices are active in each OFDM symbol. The data symbol  $x_{k,t}$  sent by the  $k$ -th device during the  $t$ -th OFDM symbol belongs to a modulation constellation set  $\Omega = \{a_1, a_2, \dots, a_L\}$ . If the device is silent in a period, then  $x_{k,t}$  is equal to zero, i.e., no signal is transmitted. To distinguish the signals of the  $k$ -th device,  $x_{k,t}$  is spread across  $M$  subcarriers through a unique spreading sequence  $\mathbf{s}_k \in \mathbb{C}^{M \times 1}$ . At the BS, the received signal  $\mathbf{y}_t \in \mathbb{C}^{M \times 1}$  can be written as

$$\mathbf{y}_t = \sum_{k=1}^K \mathbf{H}_k \mathbf{s}_k \alpha_{k,t} x_{k,t} + \mathbf{w}_t, \quad (1)$$

where  $\mathbf{H}_k = \text{diag}([h_{1,k}, h_{2,k}, \dots, h_{M,k}]^T) \in \mathbb{C}^{M \times M}$ ,  $h_{m,k}$  is the subchannel between the  $k$ -th device and the BS,  $\alpha_{k,t}$  is the binary activity indicator that is equal to one if the  $k$ -th device is active and to zero otherwise,  $\mathbf{w}_t$  is the additive white Gaussian noise (AWGN) that is distributed as  $\mathbf{w}_t \sim \mathcal{CN}(\mathbf{w}_t; \mathbf{0}_M, \sigma^2 \mathbf{I}_M)$ . We assume that  $E[|x_{k,t}|^2] = 1$  for active devices due to power control.

By defining  $\mathbf{S} = [\mathbf{H}_1 \mathbf{s}_1, \mathbf{H}_2 \mathbf{s}_2, \dots, \mathbf{H}_K \mathbf{s}_K] \in \mathbb{C}^{M \times K}$  and  $\mathbf{x}_t = [\alpha_{1,t} x_{1,t}, \alpha_{2,t} x_{2,t}, \dots, \alpha_{K,t} x_{K,t}]^T \in \mathbb{C}^{K \times 1}$ , then we have

$$\mathbf{y}_t = \mathbf{S} \mathbf{x}_t + \mathbf{w}_t. \quad (2)$$

For mMTC, the number of potential IoT devices  $K$  can be massively large but the number of radio resources for random access  $M$  is limited, i.e.,  $K \gg M$ . Therefore, it is impossible to employ traditional linear detectors, e.g., least-squares (LS) and linear minimum mean square error (LMMSE) detectors for solving  $\mathbf{x}_t$  due to the under-determined problem in (2). On the other hand, due to their sporadic traffic, the number of active devices  $K_a$  is much smaller than  $K$ , which indicates that the data vector  $\mathbf{x}_t$  is sparse [1], [4]. In view of this fact, the JADD can be formulated as the following optimization problem

$$\begin{aligned} \hat{\mathbf{x}}_t &= \arg \min_{\mathbf{x}_t} \|\mathbf{x}_t\|_0, \\ \text{s.t. } &\|\mathbf{y}_t - \mathbf{S} \mathbf{x}_t\|_2^2 \leq \varepsilon, x_{k,t} \in \{\Omega \cup 0\}, \forall k, \end{aligned} \quad (3)$$

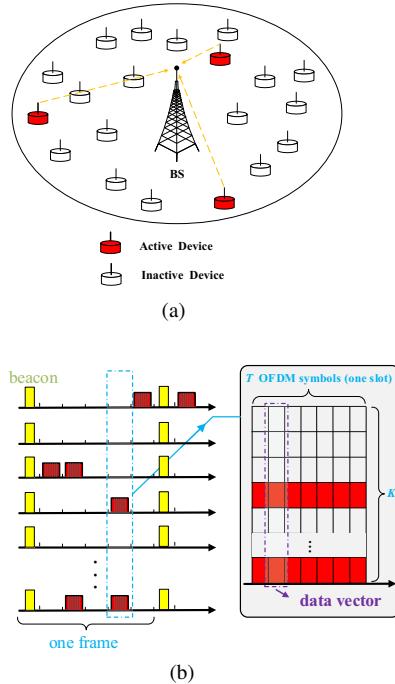


Fig. 1. (a) The activity of devices in mMTC exhibits sparsity; (b) Frame structure of the proposed beacon-aided grant-free access scheme, where the beacon is periodically broadcast by the BS to facilitate the synchronization, power control, and channel estimation at the IoT devices.

where the constant  $\varepsilon$  is associated with the noise level of the system.

We can observe that the JADD in (3) is an SMV CS problem. Therefore, sparse signal recovery algorithms can be employed to solve (3). In the next section, we will present a slotted grant-free random access protocol, where the JADD can be further formulated as an MMV CS problem for achieving better massive access performance.

### III. PROPOSED SCHEME

In this section, we propose a beacon-aided grant-free access scheme for mMTC. First, the frame structure and access procedure are presented. Then, we formulate the JADD problem as an MMV CS problem, whereby the inherently structured sparsity is exploited to design the OAMP-MMV-SSL algorithm and OAMP-MMV-ASL algorithm. Finally, SIC and channel coding are integrated into the OAMP-MMV algorithms for achieving better performance.

#### A. Frame Structure and Access Procedure

We consider that each frame starts with a beacon broadcast from the BS. The transmission of beacons is followed by the uplink random access phase, which consists of multiple time slots and each time slot consists of  $T$  OFDM symbols, as shown in Fig. 1(b). This periodic beacons can facilitate the IoT devices to achieve synchronization, power control, reliable channel estimation, etc. We consider the accurate acquisition of CSI at the device, namely  $\mathbf{H}_k \hat{\mathbf{H}}_k^{-1} = \mathbf{I}$ , where  $\hat{\mathbf{H}}_k = \text{diag}([\hat{h}_{1,k}, \hat{h}_{2,k}, \dots, \hat{h}_{M,k}]^T) \in \mathbb{C}^{M \times M}$  denotes the estimated channel. Previous works, e.g., [14]–[21], requires the CSI to be known at the BS. This indicates that extra pilot signals are necessary before the data transmission phase [9], which

increases the access latency. By contrast, the proposed scheme can fully exploit the inherent beacons that are periodically broadcast by the BS so that the random access procedure can be considerably simplified.

By performing the pre-equalization before transmitting the signals, the equivalent sensing matrix is

$$\begin{aligned}\tilde{\mathbf{S}} &= [\mathbf{H}_1 \hat{\mathbf{H}}_1^{-1} \mathbf{s}_1, \mathbf{H}_2 \hat{\mathbf{H}}_2^{-1} \mathbf{s}_2, \dots, \mathbf{H}_K \hat{\mathbf{H}}_K^{-1} \mathbf{s}_K] \\ &= [\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_K].\end{aligned}\quad (4)$$

Accordingly, the received signal at the BS can be rewritten as

$$\mathbf{y}_t = \tilde{\mathbf{S}} \mathbf{x}_t + \mathbf{w}_t. \quad (5)$$

By stacking the received signals in  $T$  OFDM symbols (i.e., one time slot), we have

$$\mathbf{Y} = \tilde{\mathbf{S}} \mathbf{X} + \mathbf{W}, \quad (6)$$

where  $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T] \in \mathbb{C}^{M \times T}$ ,  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{C}^{K \times T}$  and  $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_T]$ . As illustrated in Fig. 1(b), the IoT devices' activity remains unchanged during  $T$  OFDM symbols, i.e.,

$$\text{supp} \{\mathbf{x}_1\} = \text{supp} \{\mathbf{x}_2\} = \dots = \text{supp} \{\mathbf{x}_T\}, \quad (7)$$

where  $\text{supp}\{\cdot\}$  denotes the set of non-zero elements in a vector. Given  $\tilde{\mathbf{S}}$  and  $\mathbf{Y}$ , the aim of JADD is to estimate the sparse matrix  $\mathbf{X}$  and to identify the active devices, which results in an MMV CS problem. By leveraging the sparsity structure in (7), a more accurate identification of the active devices and improved data detection performance can be obtained.

#### B. Spreading Code

In conventional AMP algorithm, the sensing matrix is usually required to be a random matrix with i.i.d. elements, e.g., an i.i.d. Gaussian random matrix. Sensing matrices whose elements are not i.i.d. could deteriorate the performance of the AMP algorithm. On the other hand, the OAMP algorithm relaxes the sensing matrix requirement and can work well for partial unitary matrix<sup>2</sup>. Therefore, it can be applied to a wider range of applications as compared with the AMP algorithm.

In this paper, we adopt a partial discrete Fourier transformation (DFT) matrix as the sensing matrix due to the following two reasons: (a) matrix inversion can be avoided, which reduces the computational complexity; (b) a partial DFT matrix is more rhythmic than Gaussian random matrix, which is beneficial for quantization and storage. The generation of the partial DFT matrix is performed as follows. Define  $\mathbf{F} \in \mathbb{C}^{K \times K}$  as a DFT unitary matrix, i.e.,  $\mathbf{F} \mathbf{F}^H = \mathbf{I}_K$ , and  $\mathbf{P} \in \mathbb{C}^{M \times K}$  as a selection matrix generated by randomly extracting  $M$  rows from  $\mathbf{I}_K$ . Then, we have  $\tilde{\mathbf{S}} = \mathbf{P} \mathbf{F}$ .

It is worth noting that the data symbol can be spread not only in the multiple frequency-domain subcarriers, but also in the multiple time-domain chips, if narrowband systems are considered. These two mathematical formulations are equivalent from the signal processing perspective. In this paper, we

<sup>2</sup>A matrix  $\mathbf{A} \in \mathbb{C}^{M \times K}$  ( $M < K$ ) is termed as partial unitary matrix if it satisfies  $\mathbf{A} \mathbf{A}^H = \mathbf{I}_M$ .

mainly consider the former methods that is used in wideband OFDM systems (without loss of generality).

### C. OAMP-MMV Algorithm for JADD

Different from the conventional OAMP algorithm that limited to the SMV problem, we propose the OAMP-MMV-SSL algorithm and OAMP-MMV-ASL algorithm to exploit the structured sparsity in (7). Particularly, the linear estimation (LE) module and non-linear estimation (NLE) module of OAMP algorithm are denoted as module A and module B, respectively, as shown in Fig. 2. In addition, we add module C consisting of EM update. Specifically, the module B works as the data detector since it includes the minimum mean square error (MMSE) estimation of  $\mathbf{X}$ , and module C outputs the posterior sparsity ratio at the final stage for activity identification.

1) *OAMP-MMV-SSL Algorithm*: Previous works, e.g., [14]–[17], fail to fully exploit the a priori distribution of random access signals. By contrast, our proposed algorithm accounts, in the a priori distribution, for the feature that access signals originate from a discrete constellation set. Specifically, the distribution of data symbol  $x_{k,t}$  of the  $k$ -th device during the  $t$ -th OFDM symbol is assumed to be

$$p(x_{k,t}) = (1 - \lambda_{k,t})\delta(x_{k,t}) + \frac{\lambda_{k,t}}{L} \sum_{l=1}^L \delta(x_{k,t} - a_l), \quad (8)$$

where  $\lambda_{k,t} \in [0, 1]$  denotes the non-zero probability of  $x_{k,t}$  and is termed as the sparsity ratio, and  $\delta(\cdot)$  denotes the Dirac function.

In a conventional OAMP algorithm, the module A is used to decouple the received signal  $\mathbf{Y}$ . After some matrix manipulations, we have

$$\mathbf{r}_t^i = \mathbf{u}_t^{i-1} + \frac{K}{M} \tilde{\mathbf{S}}^H \left( \mathbf{y}_t - \tilde{\mathbf{S}} \mathbf{u}_t^{i-1} \right), \quad (9)$$

where  $\mathbf{r}_t^i$  is an intermediate variable that is output from module A,  $i$  denotes the  $i$ -th iteration, and  $\mathbf{u}_t$  is the NLE of  $\mathbf{x}_t$ , as described in (11). We define  $\tau_t^i = \frac{1}{K} \mathbb{E} \left[ \|\mathbf{r}_t^i - \mathbf{x}_t\|_2^2 \right]$  as the error measure of module A. On the other hand, module B is based on the assumption that  $\mathbf{r}_t$  is modeled as an AWGN observation of the real signals  $\mathbf{x}_t$ , i.e.,

$$\mathbf{r}_t = \mathbf{x}_t + \sqrt{\tau_t} \mathbf{z}_t, \quad (10)$$

where  $\mathbf{z}_t \sim \mathcal{CN}(\mathbf{z}_t; \mathbf{0}_K, \mathbf{I}_K)$  and  $\mathbf{z}_t$  is independent of  $\mathbf{x}_t$ . Then, the NLE of module B is calculated as

$$u_{k,t}^i = C_t^i \left( \mathbb{E} [x_{k,t}|r_{k,t}^i] - \frac{\bar{\gamma}_t^i}{\tau_t^i} r_{k,t}^i \right), \quad \forall k, \quad (11)$$

where  $C_t^i = \frac{\tau_t^i}{\tau_t^i - \bar{\gamma}_t^i}$  and  $\bar{\gamma}_t^i = \frac{1}{K} \sum_{k=1}^K \text{var} [x_{k,t}|r_{k,t}^i]$ . The error measure of the NLE is  $v_t^i = \frac{1}{K} \mathbb{E} \left[ \|\mathbf{u}_t^i - \mathbf{x}_t\|_2^2 \right]$ . These two modules are executed iteratively until convergence. Finally, the output of OAMP algorithm is  $\hat{x}_{k,t} = \mathbb{E} [x_{k,t}|r_{k,t}^i], \forall k$ . In addition,  $\tau_t$  and  $v_t$  are defined as [27]

$$\tau_t = \frac{K-M}{M} v_t^{i-1} + \frac{K}{M} \sigma^2, \quad (12)$$

$$v_t^i = \left( \frac{1}{\bar{\gamma}_t^i} - \frac{1}{\tau_t^i} \right)^{-1}. \quad (13)$$

Combining the observation model (10) with the a priori distribution in (8), we can obtain the approximate posterior distribution of  $x_{k,t}$ , which can be expressed as

$$p(x_{k,t}|r_{k,t}^i) = (1 - \pi_{k,t}^i)\delta(x_{k,t}) + \pi_{k,t}^i \frac{\sum_{l=1}^L q_{k,t,l}^i \delta(x_{k,t} - a_l)}{\sum_{l=1}^L q_{k,t,l}^i}, \quad (14)$$

where

$$\pi_{k,t}^i = \left[ 1 + (1 - \lambda_{k,t}) / \left( \frac{\lambda_{k,t}}{L} \sum_{l=1}^L q_{k,t,l}^i \right) \right]^{-1}, \quad (15)$$

and

$$q_{k,t,l}^i = \exp \left( - \frac{|a_l|^2 - 2 \Re \{ a_l^* r_{k,t}^i \}}{\tau_t^i} \right). \quad (16)$$

The parameter  $\pi_{k,t} \in [0, 1]$  is termed as the posterior sparsity ratio, which represents the non-zero probability of  $x_{k,t}$  based on the observation  $r_{k,t}$ . When  $\pi_{k,t} \rightarrow 0$ ,  $p(x_{k,t}|r_{k,t}^i)$  can be well approximated by a Dirac function, which implies that the  $k$ -th device is inactive in the  $t$ -th OFDM symbol. On the contrary, when  $\pi_{k,t} \rightarrow 1$ , the  $k$ -th device can be identified as an active device. Given (14), the posterior mean and variance can be formulated as

$$\mu_{k,t}^i = \mathbb{E} [x_{k,t}|r_{k,t}^i] = (\pi_{k,t}^i \sum_{l=1}^L a_l q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i), \quad (17)$$

$$\begin{aligned} \gamma_{k,t}^i &= \text{var} [x_{k,t}|r_{k,t}^i] \\ &= (\pi_{k,t}^i \sum_{l=1}^L |a_l|^2 q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i) - |\mu_{k,t}^i|^2. \end{aligned} \quad (18)$$

Equations (9)–(18) constitute the key steps of the conventional OAMP algorithm [27]. However, the conventional OAMP algorithm takes the noise variance  $\sigma^2$  and sparsity ratio  $\lambda$  as known parameters, which is not easy to obtain in general. To overcome this issue, we employ the EM algorithm [28] to learn the unknown parameters, e.g.,  $\boldsymbol{\theta} = \{\sigma^2, \lambda_{k,t}, \forall k, t\}$ . The EM algorithm consists of two steps

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{i-1}) = \mathbb{E} [\ln p(\mathbf{X}, \mathbf{Y}) | \mathbf{Y}; \boldsymbol{\theta}^{i-1}], \quad (19)$$

$$\boldsymbol{\theta}^i = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{i-1}), \quad (20)$$

where  $\mathbb{E} [(\cdot) | \mathbf{Y}; \boldsymbol{\theta}^{i-1}]$  denotes the expectation conditioned on  $\mathbf{Y}$  with parameters  $\boldsymbol{\theta}^{i-1}$ . The posterior distribution in (19) can be approximately replaced by (14) from the OAMP algorithm. However, the joint optimization of  $\boldsymbol{\theta}$  can be difficult. Therefore, we adopt the incremental EM algorithm [29], which updates one parameter in each iteration while keeping the other parameters fixed. By taking the partial derivative of (19) with respect to each element of  $\boldsymbol{\theta}$  and setting the derivatives equal to zero, we obtain the update rules of  $\boldsymbol{\theta}$  as (see Appendix A for the detailed derivation)

$$\lambda_{k,t}^i = \pi_{k,t}^i, \quad \forall k, t, \quad (21)$$

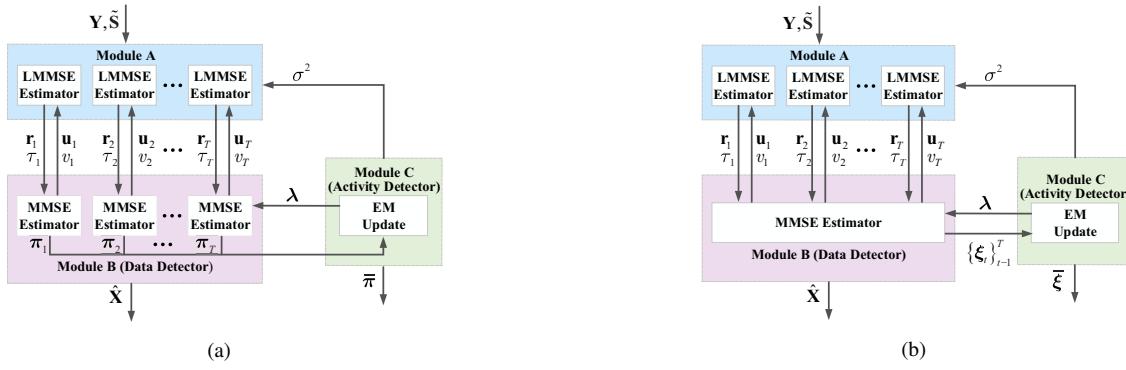


Fig. 2. Block diagram of the proposed (a) OAMP-MMV-SSL algorithm; (b) OAMP-MMV-ASL algorithm.

$$(\sigma^2)^i = \frac{1}{T} \sum_{t=1}^T \left[ \frac{1}{M} \left\| \mathbf{y}_t - \tilde{\mathbf{S}} \boldsymbol{\mu}_t^i \right\|_2^2 + \bar{\gamma}_t^i \right]. \quad (22)$$

Since the EM algorithm may converge to a local extremum or to a saddle point, a proper initialization is necessary. In our experiments, we find that the EM algorithm can work well when the initialization parameters change in a large range. Even so, the detection performance will deteriorate when the deviation between the initialization and the true value is too large. Therefore, following [28], the initialization can be tailored as

$$\lambda_{k,t}^0 = \frac{M}{K} \max_{c>0} \frac{1-2K[(1+c)^2\Phi(-c)-c\phi(c)]/M}{1+c^2-2[(1+c)^2\Phi(-c)-c\phi(c)]}, \forall k,t, \quad (23)$$

$$(\sigma^2)^0 = \frac{1}{T} \sum_{t=1}^T \frac{\|\mathbf{y}_t\|_2^2}{(\text{SNR}^0 + 1) M}, \quad (24)$$

where  $\Phi(\cdot)$  and  $\phi(\cdot)$  are the cumulative distribution function and probability distribution function of a standard Gaussian random variable, respectively, and  $\text{SNR}^0$  is set to 100 as suggested in [28]. Equation (23) sets  $\lambda_{k,t}^0$  to the theoretical phase transition point of  $\ell_1$  norm optimization based sparse signal recovery, and the second equation can be verified by substituting the received signal  $\mathbf{y}_t = \mathbf{S}\mathbf{x}_t + \mathbf{w}_t$  and the signal-noise-ratio (SNR) =  $\|\mathbf{S}\mathbf{x}_t\|_2^2/\|\mathbf{w}_t\|_2^2$  into the rightside.  $\text{SNR}^0$  is an alternative since the true SNR is unknown. The numerical results show that the proposed schemes can work well when  $\text{SNR}^0 = 100$ .

Therefore, the OAMP algorithm is capable of solving the SMV problem in (2), in which the EM steps are integrated for learning the unknown parameters  $\theta$ . If this approach is directly applied to the MMV problem in (6) and the transmitted signals  $\mathbf{X}$  are estimated column by column, however, the sparsity structure is not fully exploited. Therefore, in our proposed OAMP-MMV-SSL algorithm, we consider the following steps to learn the structured sparsity. In particular,  $\lambda_{k,t}$  represents the non-zero probability of  $x_{k,t}$  and is independently updated in (21). In view of this fact, we can refine  $\lambda_{k,t}$  as

$$\lambda_{k,1}^i = \dots = \lambda_{k,T}^i = \bar{\pi}_k^i \triangleq \frac{1}{T} \sum_{t=1}^T \pi_{k,t}^i. \quad (25)$$

Compared to (21), equation (25) learns the structured sparsi-

### Algorithm 1: OAMP-MMV-SSL Algorithm

---

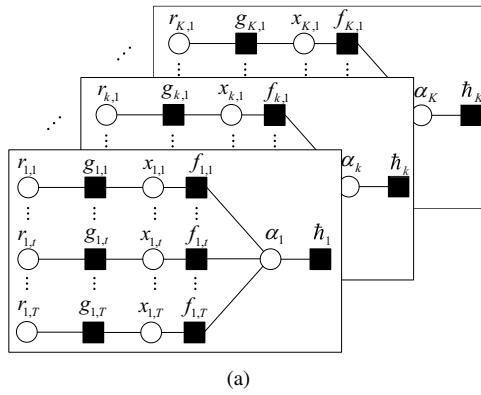
**Input:** Noisy observations  $\mathbf{Y}$ , spreading code matrix  $\tilde{\mathbf{S}}$ , and the maximum number of iterations  $I_{\text{iter}}$ .  
**Output:** Detected data matrix  $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_T]$  and the posterior sparsity ratio  $\bar{\pi}$ .

- 1:  $\forall k, t$ : Set iteration index  $i$  to 1. Initialize  $\mathbf{u}_t^0 = \mathbf{0}_K$  and  $v_t^0 = 1$ .  $\lambda_{k,t}^0$  and  $(\sigma^2)^0$  are initialized as (23) and (24), respectively.
- 2: **for**  $i = 1$  to  $I_{\text{iter}}$  **do**
- 3:   **%Module A:**
- 4:    $\forall t$ :  $\mathbf{r}_t^i = \mathbf{u}_t^{i-1} + \frac{K}{M} \tilde{\mathbf{S}}^H (\mathbf{y}_t - \tilde{\mathbf{S}} \mathbf{u}_t^{i-1})$ ,
- 5:    $\tau_t^i = \frac{K-M}{M} v_t^{i-1} + \frac{K}{M} (\sigma^2)^{i-1}$ .
- 6:   **%Module B:**
- 7:    $\forall k, t, l$ :  $q_{k,t,l}^i = \exp \left( -\frac{|a_l|^2 - 2\Re \{ a_l^* r_{k,t}^i \}}{\tau_t^i} \right)$ .
- 8:    $\forall k, t$ :  $\mu_{k,t}^i = (\pi_{k,t}^i \sum_{l=1}^L a_l q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i)$ ,
- 9:    $\gamma_{k,t}^i = (\pi_{k,t}^i \sum_{l=1}^L |a_l|^2 q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i) - |\mu_{k,t}^i|^2$ .
- 10:    $\forall t$ :  $\bar{\gamma}_t^i = \frac{1}{K} \sum_{k=1}^K \gamma_{k,t}^i$ ,    $C_t^i = \frac{\tau_t^i}{\tau_t^i - \bar{\gamma}_t^i}$ ,    $v_t^i = \left( \frac{1}{\bar{\gamma}_t^i} - \frac{1}{\tau_t^i} \right)^{-1}$ .
- 11:    $\forall k, t$ :  $u_{k,t}^i = C_t^i \left( \mu_{k,t}^i - \frac{\bar{\gamma}_t^i}{\tau_t^i} r_{k,t}^i \right)$ .
- 12:   **%Module C:**
- 13:    $\forall k, t$ :  $\lambda_{k,t}^i = \bar{\pi}_k^i = \frac{1}{T} \sum_{t=1}^T \pi_{k,t}^i$ .
- 14: **end for**
- 15:  $\forall t$ :  $\hat{\mathbf{x}}_t = \mu_t^i$ .

---

ty by averaging the posterior sparsity ratio. Then, this sparsity structure can be exploited by the MMSE estimator in the next iteration. The EM update rules in (22) and (25) constitute the module C. We summarize the proposed OAMP-MMV-SSL algorithm in Algorithm 1, whose flow chart diagram and procedure are illustrated in Fig. 2(a).

Based on the output of Algorithm 1, we can identify the active devices and detect the corresponding data. Hence, the design of the activity detector is essential to the JADD performance. Specifically, the binary hypothesis testing of the



(a)

Fig. 3. (a) Factor graph  $\mathcal{G}$  for the joint distribution  $p(\mathbf{X}, \mathbf{R}, \boldsymbol{\alpha})$  in (33); (b) The functions of the factor node in graph  $\mathcal{G}$ .

activity detection can be written as follows

$$\begin{cases} \mathcal{H}_1 : \hat{\alpha}_{k,t} = 1, \text{active}, \\ \mathcal{H}_0 : \hat{\alpha}_{k,t} = 0, \text{inactive}. \end{cases} \quad (26)$$

The corresponding two statuses are both contained in the posterior distribution in (14). By integrating over the corresponding hypothesis space, we obtain the probability of  $\mathcal{H}_1$  and  $\mathcal{H}_0$

$$\Pr(\hat{\alpha}_{k,t} = 1 | r_{k,t}) = \int_{\mathcal{H}_1} p(x_{k,t} | r_{k,t}) dx_{k,t} = \pi_{k,t}, \quad (27a)$$

$$\Pr(\hat{\alpha}_{k,t} = 0 | r_{k,t}) = 1 - \pi_{k,t}, \quad (27b)$$

respectively. Then, a Bayesian activity detector can be formulated as

$$\frac{\Pr(\hat{\alpha}_{k,t} = 1 | r_{k,t})}{\Pr(\hat{\alpha}_{k,t} = 0 | r_{k,t})} \stackrel{\mathcal{H}_1}{\gtrless} 1. \quad (28)$$

Substituting (27) into (28), we have

$$\pi_{k,t} \stackrel{\mathcal{H}_1}{\gtrless} \frac{1}{2}, \quad \forall k, t. \quad (29)$$

Since a device is either active or inactive within a slot, the final activity detector is

$$\bar{\pi}_k \stackrel{\mathcal{H}_1}{\gtrless} \frac{1}{2}, \quad \forall k. \quad (30)$$

2) *OAMP-MMV-ASL Algorithm*: In the OAMP-MMV-SSL algorithm, the a priori distributions of  $\{x_{k,t}\}_{t=1}^T$  are assumed to be mutually independent. This simplifies the derivation of the posterior distribution and the MMSE estimation. However, this assumption fails to accurately characterize the true sparsity structure of random access signals. To this end, the OAMP-MMV-ASL algorithm is proposed by fully exploiting the sparsity ratio shared by  $T$  different data symbols of the  $k$ -th device. More specifically, different from (8) that considers only the a priori distribution of  $x_{k,t}$ , the a priori distribution of  $\mathbf{x}_k = [x_{k,1}, x_{k,2}, \dots, x_{k,T}]^T$  is jointly considered as

$$\begin{aligned} p(\mathbf{x}_k | \boldsymbol{\alpha}_k) &= \prod_{t=1}^T p(x_{k,t} | \boldsymbol{\alpha}_k) \\ &= \prod_{t=1}^T \left[ (1 - \alpha_k) \delta(x_{k,t}) + \frac{\alpha_k}{L} \sum_{l=1}^L \delta(x_{k,t} - a_l) \right]. \end{aligned} \quad (31)$$

Factor	Distribution	Function
$g_{k,t}$	$p(r_{k,t}   x_{k,t})$	$\mathcal{CN}(x_{k,t}; r_{k,t}, \tau_t)$
$f_{k,t}$	$p(x_{k,t}   \boldsymbol{\alpha}_k)$	$(1 - \alpha_k) \delta(x_{k,t}) + \frac{\alpha_k}{L} \sum_{l=1}^L \delta(x_{k,t} - a_l)$
$h_k$	$p(\boldsymbol{\alpha}_k)$	$(1 - \lambda_k) \delta(\boldsymbol{\alpha}_k) + \lambda_k \delta(\boldsymbol{\alpha}_k - 1)$

(b)

Furthermore, the probability distribution of activity indicator  $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_K]^T$  can be written as

$$p(\boldsymbol{\alpha}) = \prod_{k=1}^K p(\alpha_k) = \prod_{k=1}^K [(1 - \lambda_k) \delta(\alpha_k) + \lambda_k \delta(\alpha_k - 1)], \quad (32)$$

where  $\lambda_k$  denotes the probability that  $\alpha_k = 1$ .

Based on the a priori distributions in (31)-(32) and the AWGN observation model (10), we obtain the joint probability distribution as follow

$$\begin{aligned} p(\mathbf{X}, \mathbf{R}, \boldsymbol{\alpha}) &= p(\mathbf{R} | \mathbf{X}) p(\mathbf{X} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) \\ &= \prod_{k=1}^K p(\mathbf{r}_k | \mathbf{x}_k) p(\mathbf{x}_k | \boldsymbol{\alpha}_k) p(\boldsymbol{\alpha}_k) \\ &= \prod_{k=1}^K \underbrace{p(\boldsymbol{\alpha}_k)}_{h_k} \prod_{t=1}^T \underbrace{p(r_{k,t} | x_{k,t})}_{g_{k,t}} \underbrace{p(x_{k,t} | \boldsymbol{\alpha}_k)}_{f_{k,t}}, \end{aligned} \quad (33)$$

where  $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_T] \in \mathbb{C}^{K \times T}$  and  $\mathbf{r}_k = [r_{k,1}, r_{k,2}, \dots, r_{k,T}]^T \in \mathbb{C}^{T \times 1}$ . According to (33), the corresponding factor graph  $\mathcal{G}$  is illustrated in Fig. 3(a), where the circles represent variable nodes, the squares represent factor nodes, and the meaning of the factor nodes is listed in Fig. 3(b).

The goal of the proposed OAMP-MMV-ASL algorithm is to obtain the approximate posterior distribution  $p(x_{k,t} | \mathbf{r}_k)$  for the MMSE estimation  $\hat{x}_{k,t} = E[x_{k,t} | \mathbf{r}_k]$ . To this end, we apply the sum-product algorithm [30] to the graph  $\mathcal{G}$  and obtain  $p(x_{k,t} | \mathbf{r}_k)$  through message passing. The details of the derivation are elaborated in Appendix B. In particular, the  $i$ -th iteration of the proposed OAMP-MMV-ASL algorithm can be summarized as follows. First,  $\mathbf{R}^i$  and  $\{\tau_t^i\}_{t=1}^T$  are obtained from module A. The message passing steps in module B are given in (60)-(65) (see Appendix B). On this basis, the approximate posterior distribution of  $x_{k,t}$  can be written as

$$p(x_{k,t} | \mathbf{r}_k^i) = \left( 1 - \pi'_{k,t}^i \right) \delta(x_{k,t}) + \pi'_{k,t}^i \frac{\sum_{l=1}^L q_{k,t,l}^i \delta(x_{k,t} - a_l)}{\sum_{l=1}^L q_{k,t,l}^i}, \quad (34)$$

where

$$\pi'_{k,t}^i = \left[ 1 + (1 - \xi_{k,t}^i) / \left( \frac{\xi_{k,t}^i}{L} \sum_{l=1}^L q_{k,t,l}^i \right) \right]^{-1}, \quad (35)$$

---

**Algorithm 2:** OAMP-MMV-ASL Algorithm

---

**Input:** Noisy observation  $\mathbf{Y}$ , spreading code matrix  $\tilde{\mathbf{S}}$ , and the maximum number of iterations  $I_{\text{iter}}$ .

**Output:** Detected data matrix  $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_T]$  and the updated sparsity ratio  $\xi$ .

- 1:  $\forall k, t$ : Set iteration index  $i$  to 1. Initialize  $\mathbf{u}_t^0 = \mathbf{0}_K$  and  $v_t^0 = 1$ .  $\lambda_k^0$  and  $(\sigma^2)^0$  are initialized as (23) and (24), respectively.
- 2: **for**  $i = 1$  to  $I_{\text{iter}}$  **do**
- 3:   **%Module A:**
- 4:    $\forall t$ :  $\mathbf{r}_t^i = \mathbf{u}_t^{i-1} + \frac{K}{M} \tilde{\mathbf{S}}^H (\mathbf{y}_t - \tilde{\mathbf{S}} \mathbf{u}_t^{i-1})$ ,
- 5:    $\tau_t^i = \frac{K-M}{M} v_t^{i-1} + \frac{K}{M} (\sigma^2)_t^{i-1}$ .
- 6:   **%Module B:**
- 7:    $\forall k, t, l$ :  $q_{k,t,l}^i = \exp\left(-\frac{|a_l|^2 - 2\Re\{a_l^* r_{k,t}^i\}}{\tau_t^i}\right)$ .
- 8:    $\forall k, t$ :  $\eta_{k,t}^i = 1 - \frac{1}{1 + \frac{1}{L} \sum_{l=1}^L q_{k,t,l}^i}$ ,
- 9:    $\xi_{k,t}^i = \frac{\lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}{(1 - \lambda_k^{i-1}) \prod_{t' \in \setminus t} (1 - \eta_{k,t'}^i) + \lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}$ .
- 10:    $\forall k, t$ :  $\pi_{k,t}^{\prime,i} = \left[1 + (1 - \xi_{k,t}^i) / \left(\frac{\xi_{k,t}^i}{L} \sum_{l=1}^L q_{k,t,l}^i\right)\right]^{-1}$ .
- 11:    $\forall k, t$ :  $\mu_{k,t}^i = (\pi_{k,t}^{\prime,i} \sum_{l=1}^L a_l q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i)$ ,
- 12:    $\gamma_{k,t}^i = (\pi_{k,t}^{\prime,i} \sum_{l=1}^L |a_l|^2 q_{k,t,l}^i) / (\sum_{l=1}^L q_{k,t,l}^i) - |\mu_{k,t}^i|^2$ .
- 13:    $\forall t$ :  $\bar{\gamma}_t^i = \frac{1}{K} \sum_{k=1}^K \gamma_{k,t}^i$ ,
- 14:    $C_t^i = \frac{\tau_t^i}{\tau_t^i - \bar{\gamma}_t^i}$ ,
- 15:    $v_t^i = \left(\frac{1}{\bar{\gamma}_t^i} - \frac{1}{\tau_t^i}\right)^{-1}$ .
- 16: **end for**
- 17: **forall**  $t$ ,  $\hat{\mathbf{x}}_t = \mu_t^i$ .

---

$\xi_{k,t}^i$  is given in (63), and  $\pi_{k,t}^{\prime,i}$  denotes the posterior sparsity ratio. The structure of  $p(x_{k,t}|\mathbf{r}_k)$  is similar to (14) for the OAMP-MMV-SSL algorithm. However, the calculations of  $p(x_{k,t}|\mathbf{r}_k)$  for the SSL and ASL models are different in two main aspects. As for the ASL model, the parameter  $\lambda_{k,t}$  in (15) is replaced by  $\xi_{k,t}$  in order to fully exploit the global information of  $\mathbf{r}_k$  through message passing. Also, the posterior mean  $\mu_{k,t}^i$  and variance  $\gamma_{k,t}^i$  can still be calculated in (17) and (18), but  $\pi_{k,t}$  needs to be replaced by  $\pi_{k,t}^{\prime,i}$  in (35). Moreover, similar to the OAMP-MMV-SSL algorithm, we adopt the EM algorithm to learn the unknown sparsity ratio  $\lambda_k$  and noise variance  $\sigma^2$  in module C. The update rule for  $\lambda_k$  is

$$\lambda_k^i = \bar{\xi}_k^i \triangleq \frac{1}{T} \sum_{t=1}^T \xi_{k,t}^i, \quad (36)$$

and  $(\sigma^2)^i$  is updated by employing (22). The OAMP-MMV-ASL algorithm is summarized in Algorithm 2.

In a nutshell, the main difference between the SSL and ASL models lies in the MMSE estimator (module B), as shown in Fig. 2. It originates from the different formulations for the a priori distribution of the random access matrix  $\mathbf{X}$ . As for the SSL model, in particular, we assume the mutual independence

of  $\{x_{k,t}\}_{t=1}^T$  and exploit only the sparsity structure in the EM update rules. In contrast, the a priori distribution of the ASL model is closer to true distribution and is more accurate. By resorting to the sum-product rule, the calculation of  $\xi_{k,t}$  exploits the whole information from  $\mathbf{r}_k$ . Except for the different sparsity exploitation methods, the EM update rules are the same for the ASL and SSL models. In the simulation section, we will compare the proposed OAMP-MMV-SSL and OAMP-MMV-ASL algorithms by considering different system configurations, and their advantages and limitations are discussed. We will also see that even if the system is slightly overloading (i.e.,  $M < K_a$ ), the OAMP-MMV-ASL algorithm can achieve considerable detection performance, as long as  $T$  is large enough.

#### D. SIC-Based JADD

In this section, we introduce SIC methods to enhance the performance of the proposed slotted grant-free massive access scheme. SIC methods are useful in non-linear detection and provide a trade-off between computational complexity and reliability. SIC methods are beneficial for sparse signal recovery algorithms as well [31], [32]. Taking problem (5) as an example, when the signal  $\mathbf{s}_k \alpha_{k,t} x_{k,t}$  of the active devices is detected correctly and subtracted from the received signal  $\mathbf{y}_t$ , the residual  $\tilde{\mathbf{y}}_t$  is made of fewer active devices. In particular, the residual  $\tilde{\mathbf{y}}_t$  can be written as

$$\tilde{\mathbf{y}}_t = \sum_{k=1, k \notin \kappa}^K \mathbf{s}_k \alpha_{k,t} x_{k,t} + \mathbf{w}_t, \quad (37)$$

where  $\kappa$  denotes the set of detected active devices. Therefore, the data vector becomes sparser after performing interference cancellation. After multiple SIC operations, the number of residual active devices becomes small, leading to improved detection performance.

However, reaping these advantages by using SIC methods faces several challenges, which include increasing the computational complexity, the associated error propagation, and the choice of criterion of cancellation, etc. Specifically, SIC-based detection algorithms are executed repeatedly, which dramatically increases the computational complexity. An intuitive solution to this problem is to limit the number of interference cancellations, which motivates us to subtract the signals of multiple active devices simultaneously instead of one active device at each iteration. In this way, we can strive a balance between the system performance and the computational complexity. In addition, error propagation originates from the cumulative effects of the subtracted signals. To mitigate this error, we integrate channel coding into the SIC scheme, which increases the reliability of data detection thanks to the error correction capability of channel coding.

Furthermore, the cancellation ordering of SIC impacts the detection performance. This issue is usually tackled by subtracting the strongest signals first. To this end, the SNR or the signal-to-interference-and-noise-ratio (SINR) is usually used as the metrics for performing the ordering [33]. However, due to power control with  $E[|x_{k,t}|^2] = 1$ , the SNRs associated with different devices are almost the same. Also, the SINR is

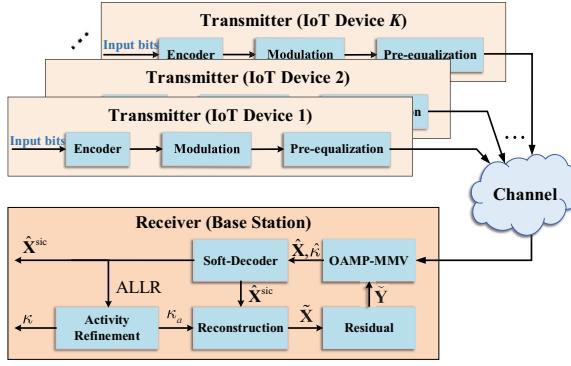


Fig. 4. Block diagram of the proposed SIC-based JADD.

usually difficult to calculate in practice. Hence, we consider the LLR obtained in the soft decision. More specifically, the LLR is the logarithm of the ratio of the probability of transmitting a 0 bit versus the probability of transmitting a 1 bit for a received signal. For a specific bit  $b$ , the LLR is defined as

$$\begin{aligned} \Lambda_{k,t,b} &= \ln \left[ \frac{\Pr(b=0|\hat{x}_{k,t})}{\Pr(b=1|\hat{x}_{k,t})} \right] \\ &= \ln \left\{ \frac{\sum_{a \in \Omega_0} \exp \left\{ -\frac{1}{\sigma^2} \left[ (x_0 - a_x)^2 + (y_0 - a_y)^2 \right] \right\}}{\sum_{a \in \Omega_1} \exp \left\{ -\frac{1}{\sigma^2} \left[ (x_0 - a_x)^2 + (y_0 - a_y)^2 \right] \right\}} \right\}, \end{aligned} \quad (38)$$

where  $b$  is the  $b$ -th bit position of  $\hat{x}_{k,t}$ ,  $\Omega_0$  and  $\Omega_1$  are the sets of ideal constellation symbols with bit 0 and 1 in the  $b$ -th bit position, respectively,  $(x_0, y_0)$  and  $(a_x, a_y)$  are the coordinates of  $\hat{x}_{k,t}$  and the ideal symbol  $a$  in the constellation diagram, respectively. The noise variance  $\sigma^2$  in (38) can be obtained from module C of our proposed OAMP-MMV algorithm. Therefore, the signals whose module of LLR is larger are more likely to be correctly detected and decoded. Although the exact LLR can provide the greatest accuracy, obtaining such information is computationally expensive. Hence, as a compromise approach, we consider to employ the approximate LLR (ALLR), by trading the accuracy for higher efficiency, that is defined as

$$\begin{aligned} \Lambda_{k,t,b} &\approx -\frac{1}{\sigma^2} \left\{ \min_{a \in \Omega_0} \left[ (x_0 - a_x)^2 + (y_0 - a_y)^2 \right] \right. \\ &\quad \left. - \min_{a \in \Omega_1} \left[ (x_0 - a_x)^2 + (y_0 - a_y)^2 \right] \right\}. \end{aligned} \quad (39)$$

To evaluate the reliability of the detected signal associated with the  $k$ -th device, we define the average ALLR metric as

$$\bar{\Lambda}_k = \frac{1}{T \log_2 L} \sum_{t=1}^T \sum_{b=1}^{\log_2 L} |\Lambda_{k,t,b}|. \quad (40)$$

The block diagram of the proposed SIC-based OAMP-MMV-SSL and OAMP-MMV-ASL algorithms is illustrated in Fig. 4, and the associated pseudocode is given in Algorithm 3. More specifically,  $\hat{\mathbf{X}}$  and  $\hat{\kappa} = \{k | \hat{\alpha}_k = 1, \forall k\}$  are obtained from the OAMP-MMV-SSL or OAMP-MMV-ASL algorithm as an initial estimate. Then, in line 4, the average ALLR is calculated for activity refinement. By choosing the first  $N^{\text{sic}}$

### Algorithm 3: SIC-Based OAMP-MMV Algorithm

---

**Input:** Noisy observation  $\mathbf{Y}$ , spreading code matrix  $\tilde{\mathbf{S}}$ , the number of devices in each interference cancellation  $N^{\text{sic}}$ , and the maximum number of SIC iterations  $I_{\max}$ .

**Output:** Decoded data bit matrix  $\hat{\mathbf{X}}^{\text{sic}}$  and the set of the detected active devices  $\kappa$ .

```

1: Set  $f^{\text{sic}} = 1$ ,  $i^{\text{sic}} = 0$ ,  $\tilde{\mathbf{Y}} = \mathbf{Y}$ ,  $\hat{\mathbf{X}}^{\text{sic}} = \mathbf{O}$ , and  $\kappa = \emptyset$ .
2: while  $f^{\text{sic}}$  do
3:   Obtain  $\hat{\mathbf{X}}$  and  $\hat{\kappa}$  from OAMP-MMV-SSL algorithm or
      OAMP-MMV-ASL algorithm.
4:   Calculate  $\{\bar{\Lambda}_k\}_{k \in \hat{\kappa}}$  according to (40), and then sort them in
      the descent order.
5:   if  $\text{card}\{\hat{\kappa}\} \leq N^{\text{sic}}$  then
6:      $\kappa_a = \hat{\kappa}$ ,  $f^{\text{sic}} = 0$ .
7:   else
8:     Choose the first  $N^{\text{sic}}$  largest entries of  $\{\bar{\Lambda}_k\}_{k \in \hat{\kappa}}$ , and
      these devices forms the set  $\kappa_a$ .
9:   end if
10:   $k \in \kappa_a$ ,  $\hat{\mathbf{X}}_{k,:}^{\text{sic}} \leftarrow \Xi \hat{\mathbf{X}}_{k,:}$ .
11:   $\kappa = \kappa \cup \kappa_a$ .
12:   $k \in \kappa_a$ ,  $\tilde{\mathbf{X}}_{k,:} \leftarrow \Xi^{-1} \hat{\mathbf{X}}_{k,:}^{\text{sic}}$ .
13:   $\tilde{\mathbf{Y}} = \mathbf{Y} - \sum_{k \in \kappa_a} \tilde{\mathbf{S}}_{:,k} \tilde{\mathbf{X}}_{k,:}$ .
14:  if  $i^{\text{sic}} > I_{\max}$  then
15:     $f^{\text{sic}} = 0$ .
16:  else
17:     $i^{\text{sic}} = i^{\text{sic}} + 1$ .
18:  end if
19: end while

```

---

largest entries of  $\{\bar{\Lambda}_k\}_{k \in \hat{\kappa}}$ , we obtain a subset  $\kappa_a$  of  $\hat{\kappa}$ . The decoded data of the devices in  $\kappa_a$  is considered to be more accurate compared with that of other devices in  $\hat{\kappa}$ . Next, as described in lines 10-11 of Algorithm 3, we obtain the decoded data and indices of the devices in  $\kappa_a$ , where  $\hat{\mathbf{X}}^{\text{sic}}$  is the final decoded data,  $\Xi$  denotes the operation of demodulation in soft decision and channel decoding. Before performing interference cancellation, the reconstructed transmitted signals  $\tilde{\mathbf{X}}$  is obtained in line 12, where  $\Xi^{-1}$  denotes the operation of channel coding and modulation. Finally, the residual  $\tilde{\mathbf{Y}}$  is obtained in line 13. When the cardinality of  $\kappa_a$  is smaller than  $N^{\text{sic}}$  or the number of SIC iterations reaches a predetermined level  $I_{\max}$ , the SIC-based OAMP-MMV-SSL algorithm or the OAMP-MMV-ASL algorithm are terminated.

## IV. PERFORMANCE ANALYSIS

In this section, the SE and the computational complexity of the proposed two algorithms are analyzed. The SE of the proposed OAMP-MMV-SSL algorithm and OAMP-MMV-ASL algorithm is derived for theoretically predicting the mean square error (MSE) and bit error rate (BER) performance. Also, the computational complexity of the proposed algorithms is compared against that of the state-of-the-art benchmarks.

### A. State Evolution

The performance of AMP-type algorithms can be analyzed by studying the SE in the large system limit, as discussed in the previous works [24], [27]. Therefore, we employ the SE to theoretically predict the performance and to optimize the associated parameters for achieving the desired performance.

---

**Algorithm 4:** SE of OAMP-MMV Algorithms

---

**Input:** The number of subcarriers  $M$ , the number of potential devices  $K$ , the maximum number of iterations  $I_{\text{iter}}$ , the number of Monte Carlo simulations  $N$ , and the noise variance  $\sigma^2$ .  
**Output:** The predicted MSE  $\vartheta^i$ .

- 1:  $\forall k, t, n$ : Set iteration index  $i$  to 1. Initialize  $v_n^0 = 1$ .  $\lambda_{k,t}^0$  is initialized as (23).
- 2: Generate Monte Carlo samples  $\{\mathbf{X}_n\}_{n=1}^N$ . Each  $\mathbf{X}_n$  follows the true a priori distribution and the sparsity structure.
- 3: Generate standard complex Gaussian noise  $\{\mathbf{z}_{t,n}\}_{t,n=1}^{T,N}$ .
- 4: **for**  $i = 1$  to  $I_{\text{iter}}$  **do**
- 5:    $\forall n, \tau_n^i = \frac{K-M}{M} v_n^{i-1} + \frac{K}{M} \sigma^2$ .
- 6:    $\forall t, n, \mathbf{r}_{t,n}^i = \mathbf{x}_{t,n} + \sqrt{\tau_n^i} \mathbf{z}_{t,n}$ .
- 7:
- 8:   **%for OAMP-MMV-SSL algorithm:**
- 9:   Perform lines 6-12 of Algorithm 1.
- 10:
- 11:   **%for OAMP-MMV-ASL algorithm:**
- 12:   Perform lines 6-13 of Algorithm 2.
- 13:
- 14:    $\vartheta^i = \frac{1}{KTN} \sum_{K=1}^K \sum_{t=1}^T \sum_{n=1}^N |\mu_{k,t,n}^i - x_{k,t,n}^i|^2$ .
- 15:    $\forall n, v_n^i = \frac{1}{KT} \sum_{t=1}^T \sum_{k=1}^K |u_{k,t,n}^i - x_{k,t,n}^i|^2$ .
- 16: **end for**

---

As far as conventional OAMP algorithms are concerned, the associated SE recursion is defined as follows

$$\text{LE : } \tau^i = \frac{K-M}{M} v^{i-1} + \frac{K}{M} \sigma^2, \quad (41a)$$

$$\text{NLE : } v^i = \mathbb{E} [|U - X|^2], \quad (41b)$$

$$\vartheta^i = \mathbb{E} \left[ \left| \mathbb{E} [X | X + \sqrt{\tau^i} Z] - X \right|^2 \right], \quad (41c)$$

where  $U = \frac{\tau^i}{\tau^i - \vartheta^i} \left\{ \mathbb{E} [X | X + \sqrt{\tau^i} Z] - \frac{\vartheta^i}{\tau^i} (X + \sqrt{\tau^i} Z) \right\}$  and  $\vartheta$  is the predicted MSE,  $X \sim p(x)$  where  $p(x)$  is the same as the a priori distribution in (8), and  $Z \sim \mathcal{CN}(z; 0, 1)$ . According to the definition of expectation,  $v^i$  and  $\vartheta^i$  can be written as

$$v^i = \iint \mathbb{E} [|U - X|^2] \mathcal{D}x \mathcal{D}z, \quad (42a)$$

$$\vartheta^i = \iint \left| \mathbb{E} [X | X + \sqrt{\tau^i} Z] - X \right|^2 \mathcal{D}x \mathcal{D}z, \quad (42b)$$

respectively, where  $\mathcal{D}x = p(x)dx$  and  $\mathcal{D}z = \exp(-|z|^2)/\pi dz$ .

Limited to the SMV problem,  $p(x)$  cannot characterize the sparsity structure in (7), which would deteriorate the prediction accuracy of SE. To take the sparsity structure into account and to simplify the calculation of (42), we adopt Monte Carlo methods to carry out SE. Particularly, as for the SE of the OAMP-MMV-SSL algorithm, we generate enough samples of  $\mathbf{X}$  according to the a priori distribution (8) and the sparsity structure (7). Based on the assumption (10), we add the Gaussian noise  $\mathbf{Z}$  to  $\mathbf{X}$  for obtaining  $\mathbf{R}$ . Then, we have the approximation

$$v_n^i \approx \frac{1}{KT} \sum_{t=1}^T \sum_{k=1}^K |u_{k,t,n}^i - x_{k,t,n}^i|^2, \quad (43a)$$

$$\vartheta^i \approx \frac{1}{KTN} \sum_{n=1}^N \sum_{t=1}^T \sum_{k=1}^K |\mu_{k,t,n}^i - x_{k,t,n}^i|^2, \quad (43b)$$

where  $N$  is the number of iterations. As for the OAMP-MMV-ASL algorithm, we adopt the same approach to make sure that the structured sparsity learning via message passing is taken into account. The whole SE recursions for OAMP-MMV-SSL and OAMP-MMV-ASL algorithms are summarized in Algorithm 4. The accuracy of the adopted approximations will be verified in the simulations section.

*Remark:* We assume the availability of the true noise variance  $\sigma^2$  for SE prediction. Numerical results show that this assumption has a negligible impact on the capacity of the prediction performance of SE. Based on the SE recursion, the BER performance can be predicted by comparing the transmitted signal  $\mathbf{X}$  and the MMSE estimate  $\hat{\mathbf{X}}$ . Numerical results demonstrate that the predicted BER performance is close to the simulated BER.

### B. Computational Complexity Analysis

In Table I, the computational complexity of the proposed algorithms is analyzed and compared with that of state-of-the-art CS baseline algorithms. The baseline algorithms include the generalized subspace pursuit (GSP) algorithm [34], the simultaneous weighted orthogonal matching pursuit (SWOMP) algorithm [35], the sparsity adaptive matching pursuit (SAMP) algorithm [36], and the AMP-MMV algorithm [37].

We choose the number of complex-valued multiplications as the metric of interest, and the complexity of one real-valued multiplication is assumed to be equal to one quarter of the complexity of one complex-valued multiplication. From Table I, we evince that the GSP, SWOMP, and SAMP algorithms (which are greedy-type CS algorithms), have the same order of computational complexity, i.e., the cube of  $K_a$ , due to the matrix inversion needed for LS estimation. Such a high computational complexity cannot accommodate the requirements of low-delay and low processing complexity for application to future mMTC systems.

By contrast, the computational complexity of the proposed OAMP-MMV algorithms increases only linearly with  $K$  and  $M$ . This is because the matrix inversion is avoided when  $\tilde{\mathbf{S}}$  is a partial DFT matrix. The computational complexity of the OAMP-MMV-ASL algorithm is higher than that of the OAMP-MMV-SSL algorithm, especially when  $T$  is considerably large. Since the data packets in mMTC are usually small, the value of  $T$  is usually small, and the gap in computational complexity is negligible. Simulation results illustrated next show that the higher computational complexity of the OAMP-MMV-ASL algorithm is worthy, since a considerable performance gain compared with the OAMP-MMV-SSL algorithm is obtained, especially when  $T$  is small.

The two SIC-based OAMP-MMV algorithms are not compared in Table I due to the following reasons. During each iteration of interference cancellation, the complete execution of the OAMP-MMV-SSL or OAMP-MMV-ASL algorithms is needed for initial estimation, as shown in line 3 of Algorithm 3. The remaining manipulations such as subtractions and demodulations give a minor contribution to the computational

TABLE I: Complexity Analysis

Algorithm	Number of complex-valued multiplications in each iteration
OAMP-MMV-SSL	$3MKT + MT + \frac{13}{4}KTL + \frac{25}{4}KT$
OAMP-MMV-ASL	$3MKT + MT + \frac{13}{4}KTL + 3KT^2 + \frac{27}{4}KT$
GSP	$(2M + 1)KT + MT^2 + 2M(T + 1)K_a^2 + 2K_a^3$
SWOMP	$MKT + MT^2 + 2MTi + 2Mi^2 + i^3$
SAMP	$(2M + 3)KT + 2M(T + 1)j^2 + T + 2j^3$
AMP-MMV	$4MKT + \frac{7}{4}MK + \frac{19}{2}KT$

Note:  $i$  denotes the iteration index [35] and  $j$  denotes the stage index [36].

complexity. In view of this fact, we choose an appropriate  $N^{sic}$  to guarantee a small number of iterations in SIC and to reduce computational complexity. Therefore, the SIC-based OAMP-MMV algorithms have about one order of magnitude higher computational complexity than the original algorithms without SIC.

## V. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed schemes for JADD in grant-free massive access. For the presented simulation results, we set  $K = 500$  potential devices in total, with  $K_a = 50$  of them being active in  $T$  OFDM symbol intervals (i.e., one time slot). The spreading codes are chosen from a partial DFT matrix and preassigned to different devices in advance. The length of the spreading code is equal to  $M$ , which is also equal to the number of subcarriers for random access and to the number of measurements in the formulated CS problem. We adopt quadrature phase shift keying (QPSK) modulation for all simulations and Turbo codes<sup>3</sup> for channel coding in Figs. 10-11. As for the parameters of the SIC-based OAMP-MMV algorithms, we set the number of devices in each interference cancellation iteration  $N^{sic}$  multiplied by the maximum iterations  $I_{\max}$ , i.e.,  $N^{sic} \cdot I_{\max}$ , to 100<sup>4</sup>. We define the activity detection error probability (AEP) and the BER as follows

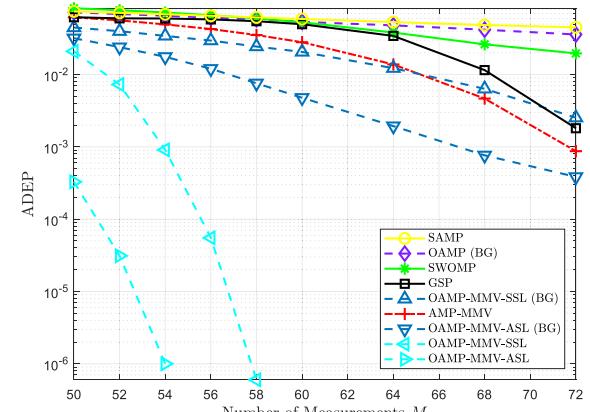
$$AEP = \frac{1}{K} \sum_{k=1}^K |\hat{\alpha}_k - \alpha_k|, \quad BER = 1 - \frac{N_s}{K_a T \log_2 L}, \quad (44)$$

where  $N_s$  denotes the number of correctly decoded bits for the successfully detected devices. The number of iterations of the OAMP-MMV algorithms  $I_{\text{iter}}$  is set to 50. Moreover, to compare the simulation results with SE-based results, we adopt the MSE metric, which is defined as

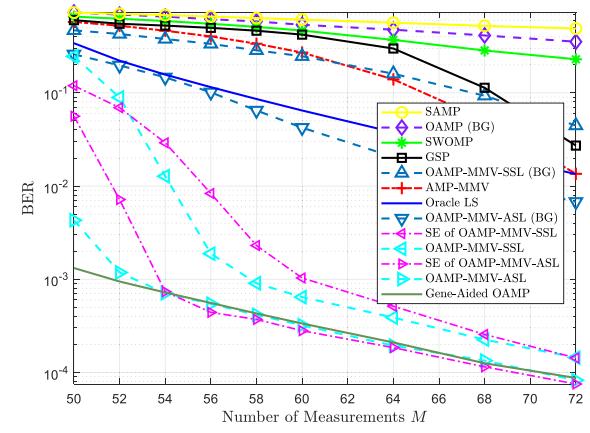
$$MSE = \frac{1}{KT} \sum_{t=1}^T E \left[ \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|_2^2 \right], \quad (45)$$

<sup>3</sup>The error correction capability of channel coding can reduce the error propagation of the proposed SIC-based schemes. Usually, the Turbo codes have better performance than the LDPC codes in the case of short code length and low code rate, while short-packet communication is one of the typical characteristics of mMTC [38]. Therefore, we adopt the Turbo codes in our proposed schemes.

<sup>4</sup>Although the exact number of active devices  $K_a$  is unknown, we assume that the BS has known that  $K_a$  is smaller than a threshold. Therefore, we set the product of  $N^{sic}$  and  $I_{\max}$ , i.e., the possible maximum number of detected active devices, to 100, which is large enough to be the threshold in our simulations.



(a)

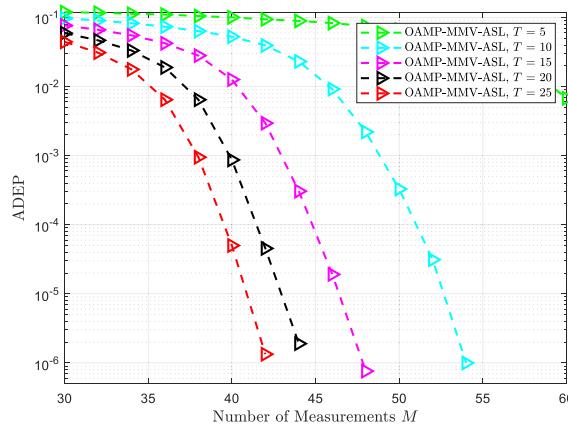


(b)

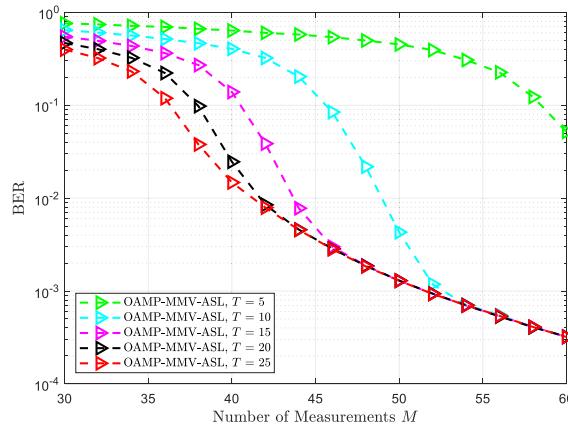
Fig. 5. Comparison of different detection algorithms<sup>5</sup> versus  $M$  given  $T = 10$  and SNR = 10 dB: (a) AEP; (b) BER.

and set  $N = 1000$ . The baseline algorithms include the GSP algorithm, SWOMP algorithm, SAMP algorithm, Oracle LS algorithm [17], AMP-MMV algorithm, and Gene-Aided OAMP algorithm, where the AMP-MMV algorithm applies (25) to leverage the sparsity structure. The Gene-Aided OAMP algorithm is the proposed OAMP-MMV algorithm

<sup>5</sup>In Fig. 5, we substitute  $v_t^i = \frac{\|\mathbf{y}_t - \mathbf{F}\mathbf{u}_t^i\|_2^2 - M\sigma^2}{M}$  of [27, (30)] into the variance  $\tau_t^i = \frac{K-M}{M}v_t^{i-1} + \frac{K}{M}(\sigma^2)^{i-1}$  in line 4 of Algorithm 2 for better detection performance of the proposed OAMP-MMV-ASL algorithm (mainly for the case of low number of observations, i.e.,  $M < 56$ , which leads to the gap between the SE and the simulation results).



(a)

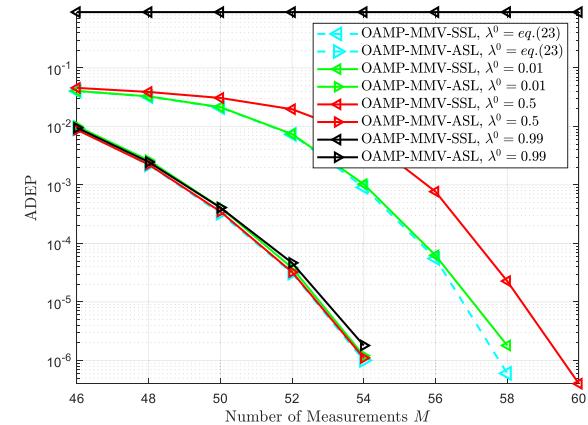


(b)

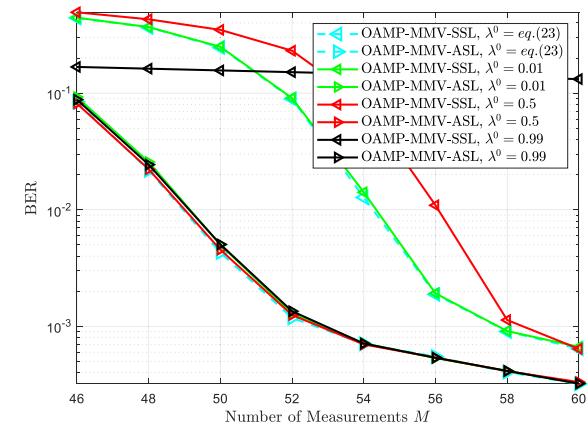
Fig. 6. Comparison of the OAMP-MMV-ASL algorithm versus  $M$  under different  $T$ , given SNR = 10 dB: (a) ADEP; (b) BER.

knowing perfect support and true noise variance. Specifically, in each iteration of the Gene-Aided OAMP algorithm, we have  $\lambda_{k,t} = \pi_{k,t} = 1$  for active devices and  $\lambda_{k,t} = \pi_{k,t} = 0$  for inactive devices. Therefore, it is the low-bound of our proposed OAMP-MMV-SSL/ASL algorithms. The Matlab codes of this paper will be released on the corresponding author's homepage <https://gaozhen16.github.io/#Latest> in the near future.

Fig. 5 compares the ADEP and BER performance of different schemes versus the spreading code length  $M$ , where the conventional OAMP algorithm and the proposed OAMP-MMV algorithm that employs a Bernoulli-Gaussian (BG) a priori distribution are considered. Compared with other algorithms, the OAMP-MMV-SSL and OAMP-MMV-ASL algorithms provide better performance. By contrast, the performance of the two proposed OAMP-MMV algorithms based on the BG a priori distribution is worse, since the prior knowledge of discrete modulation symbols is not fully exploited. Even though, it can be seen that they outperform the conventional OAMP algorithm, which owes to the utilization of sparsity structure. We observe a performance gap between the OAMP-MMV-SSL and OAMP-MMV-ASL algorithms, especially when  $M$  is small. This is because the latter algorithm can learn the structured sparsity more accurately, at the cost of a higher computational complexity. Besides, it can be observed



(a)

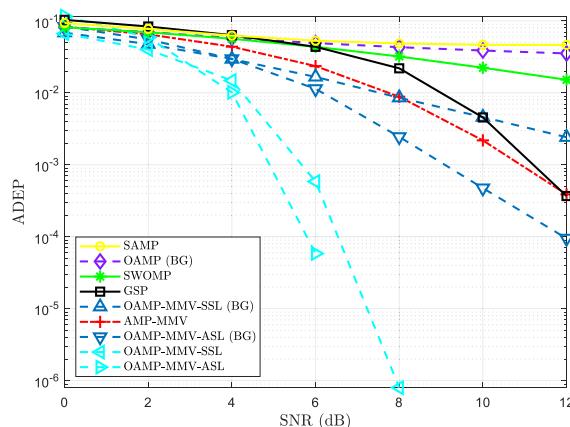


(b)

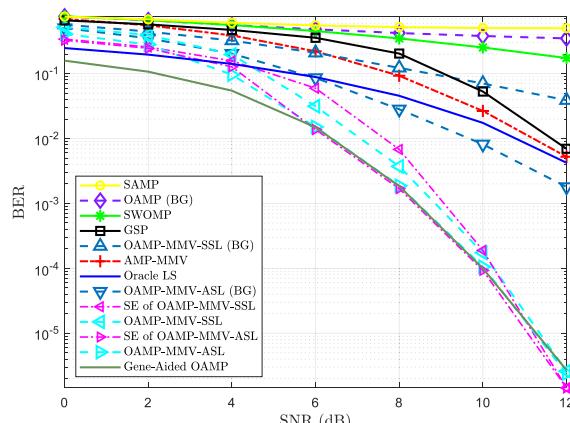
Fig. 7. Given SNR = 10 dB and  $T = 10$ , the detection performance of the OAMP-MMV-SSL/ASL algorithms versus  $M$  with different initializations of the sparsity ratio  $\lambda^0$ : (a) ADEP; (b) BER.

that when  $M$  is larger than 52, the BER performance of the proposed OAMP-MMV-ASL algorithm is almost overlapped with that of the Gene-Aided OAMP algorithm. When the number of observation is large enough, i.e.,  $M > 58$ , the two proposed OAMP-MMV algorithms can detect the active device with a high probability of success, and their BER gap becomes small. The SE results of the two proposed OAMP-MMV algorithms are presented in Fig. 5(b), which shows that the BER performance is well predicted when  $M$  is larger than 60. Moreover, the baseline algorithms need more observations to obtain good performance. The Oracle LS that assumes the perfect activity detection is the low-bound of the greedy algorithms above. However, it is still inferior to the proposed OAMP-MMV-SSL/ASL algorithms in terms of BER performance, since it cannot utilize the prior information of signals' discrete constellation modulation in the stage of data detection.

Given SNR = 10 dB, Fig. 6 compares the performance of the proposed OAMP-MMV-ASL algorithm under a small number of measurements with different  $T$ . It is worth noting that the OAMP-MMV-ASL algorithm can obtain good detection performance even if it is overloading (i.e.,  $M < K_a$ ), as long as  $T$  is large enough. For example, when  $T = 25$ ,



(a)

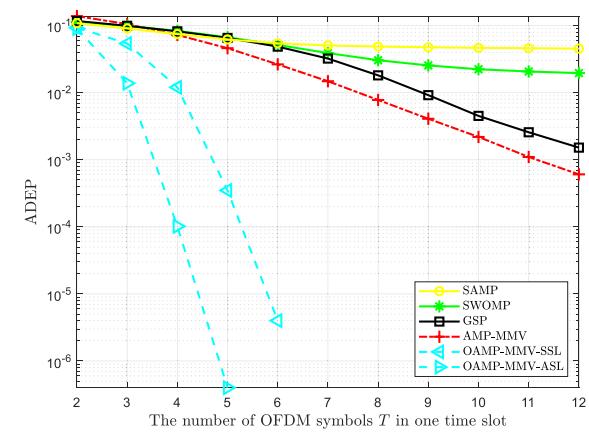


(b)

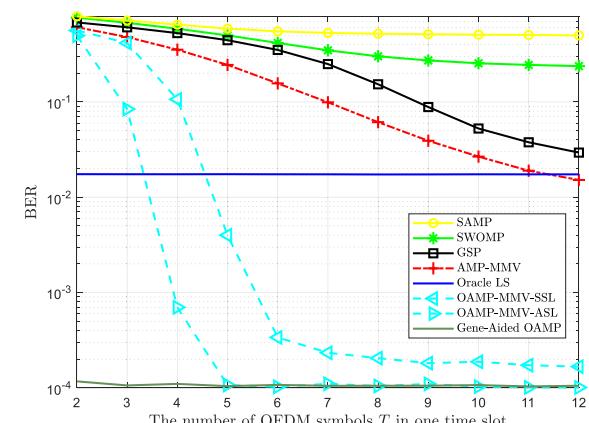
Fig. 8. Comparison of different detection algorithms versus SNR given  $T = 10$  and  $M = 70$ : (a) ADEP; (b) BER.

the proposed scheme can achieve a good ADEP of  $10^{-4}$  and BER of  $10^{-2}$  under the overloading rate  $K_a/M = 125\%$ . This result can be further improved by adding channel coding and SIC processing into the proposed scheme. Besides, it can be seen that with the increase of  $T$ , the OAMP-MMV-ASL algorithm arriving at a high probability of successful detection has a lower requirement on  $M$ . Particularly, if we define the high probability of success is  $\text{AEP} > 10^{-6}$ , the required minimum number of measurements  $M_{\min}$  is 43, 45, and 48 when  $T$  is equal to 25, 20, and 15, respectively, given SNR = 10 dB.

The performance comparison of the proposed schemes under different initializations of EM parameters is provided in Fig. 7, given SNR = 10 dB and  $T = 10$ . Both the proposed algorithms are robust to the initialization of noise variance ( $\sigma^2$ )<sup>0</sup>, so we only provide the numerical results with different  $\lambda^0$ . For interesting readers, we recommend the complete comparison on the arXiv website (<https://arxiv.org/abs/2011.07928>). In Fig. 7,  $(\sigma^2)^0$  is initialized as equation (24), and  $\lambda^0$  is set to 0.01, 0.5, 0.99, and according to equation (23), respectively. Among these four settings, we can see that the last one has the best detection performance for the OAMP-MMV-SSL algorithm.  $\lambda^0 = 0.5$  represents that there is no prior information about the sparsity, leading to the



(a)

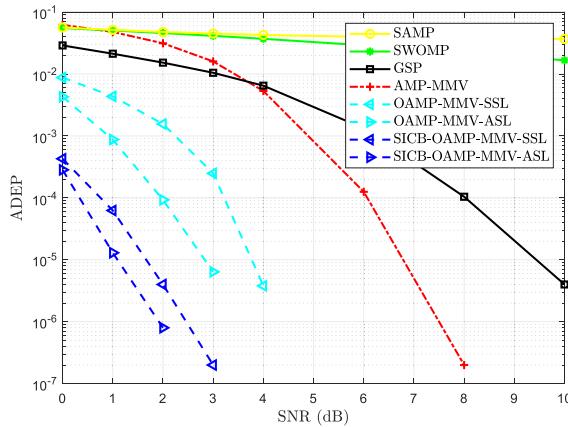


(b)

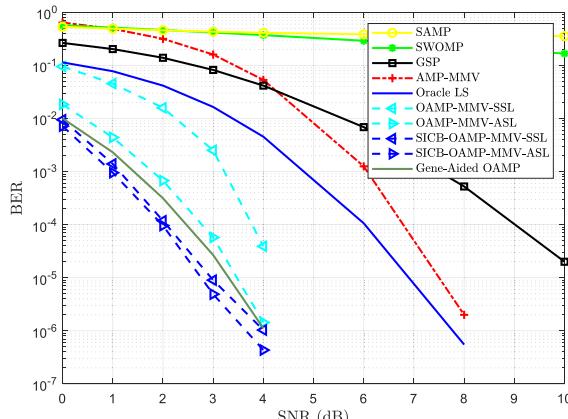
Fig. 9. Comparison of different detection algorithms versus  $T$  given  $M = 70$  and SNR = 10 dB: (a) ADEP; (b) BER.

performance deterioration of the OAMP-MMV-SSL algorithm. When  $\lambda^0$  increases to 0.99, it does not work at all. On the contrary, the OAMP-MMV-ASL algorithm works well, no matter how  $\lambda^0$  changes. These results indicate that for the OAMP-MMV-SSL algorithm, the EM method can perform well in estimating the sparsity ratio unless its initialization is extremely deviated. The OAMP-MMV-ASL algorithm is more robust to the change of  $\lambda^0$ , since it has a stronger capability of learning the sparsity structure than the OAMP-MMV-SSL method.

Fig. 8 compares the ADEP and BER performance of different schemes as a function of the SNR. It can be observed that, when the SNR is large, both ADEP and BER of our proposed OAMP-MMV algorithms decrease rapidly and outperform other algorithms significantly. The performance of the OAMP-MMV-SSL algorithm is close to that of the OAMP-MMV-ASL algorithm when the SNR is larger than 8 dB, since the sparsity structure in the high SNR range can be well learned. In addition, Fig. 8(b) demonstrates that the simulated BER performance can be predicted by the SE results in the high SNR regime. It is worth noting that the proposed OAMP-MMV-ASL algorithm has the same BER performance as that of the Gene-Aided OAMP algorithm when SNR > 6 dB. However, the greedy algorithms have poor performance even



(a)

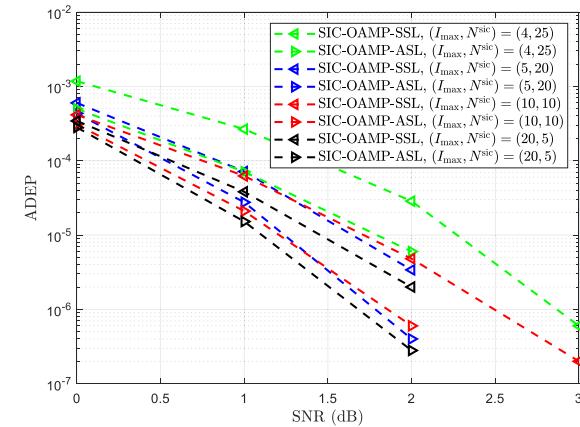


(b)

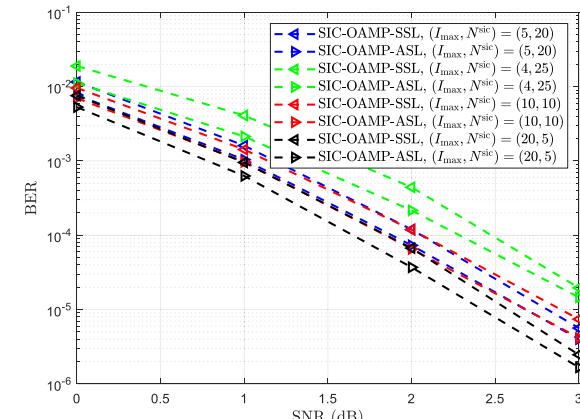
Fig. 10. Gains of the proposed SIC-based OAMP-MMV algorithms versus SNR given  $M = 70$ ,  $T = 51$ : (a) ADEM; (b) BER.

when the SNR is relatively high, as they fail to fully exploit the prior knowledge of transmitted signals and the sparsity structure.

Fig. 9 compares the ADEM and BER performance of the considered schemes as a function of the number  $T$  of OFDM symbols in each time slot. All algorithms perform poorly when  $T = 2$ . This is because the MMV model boils down to the SMV model when  $T$  is close to 1, resulting in limited gains when exploiting the sparsity structured in (7). As  $T$  increases, the OAMP-MMV-ASL and OAMP-MMV-SSL algorithms distinctively outperform other baseline algorithms. The data detection performance of the Oracle LS does not improve with the increase of  $T$ . Though it knows all active devices, its BER performance is about two orders of magnitude worse than our proposed schemes when  $T > 8$ . In addition, Fig. 9 demonstrates that the OAMP-MMV-ASL algorithm can learn the sparsity structure more accurately than the OAMP-MMV-SSL algorithm, especially when  $T$  is small. In particular, when  $T$  is larger than 7, the two proposed OAMP-MMV algorithms can achieve almost perfect activity identification, as shown in Fig. 9(a), which verifies the significant gain benefited from exploiting the sparsity structure. This observation indicates that it is better to adopt the OAMP-MMV-SSL algorithm when  $T$  is large, which has lower computational complexity and



(a)



(b)

Fig. 11. The comparison of detection performance of the SIC-based OAMP-MMV algorithms versus SNR with different pairs of  $N^{sic}$  and  $I_{max}$ , given  $M = 70$  and  $T = 51$ : (a) ADEM; (b) BER.

negligible BER performance loss compared to the OAMP-MMV-ASL algorithm. In contrast, the OAMP-MMV-ASL algorithm is more suitable if  $T$  is small. A BER floor of the proposed algorithms can be observed in Fig. 9(b). This is because the device's activity can be almost perfectly obtained if  $T > 7$ , while the fixed SNR is the key limiting factor for BER improvement.

Fig. 10 shows the effectiveness of the two proposed SIC-based OAMP-MMV algorithms with  $N^{sic} = I_{max} = 10$ , where all schemes consider Turbo coding. In particular, a code rate equal to 1/3, a code length of 30, and a number of tail bits of 12 are considered. It can be observed that the proposed SIC-based schemes achieve better ADEM and BER performance than the OAMP-MMV-SSL and OAMP-MMV-ASL algorithms. The two proposed SIC-based schemes have similar BER performance, even though their ADEM gap is still apparent. Therefore, we conclude that the SIC-based OAMP-MMV-SSL algorithm is more practical considering the offered trade-off in terms of performance and computational complexity.

In order to investigate the performance difference of the proposed SIC-based schemes under different  $N^{sic}$ , we provide the ADEM and BER performance in Fig. 11, where

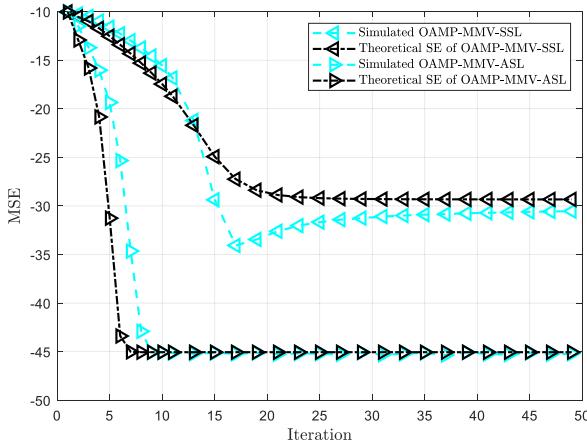


Fig. 12. MSE comparison between the simulated results and theoretical SE, given  $M = 70$ , SNR = 10 dB, and  $T = 10$ .

$N^{\text{sic}} \cdot I_{\max} = 100$  is considered. It can be observed that the detection performance of the proposed SIC-based OAMP-MMV-SSL/ASL algorithms under  $(I_{\max}, N^{\text{sic}}) = (20, 5)$  is superior to that under other conditions, and the worst performance is obtained under the case of  $I_{\max} = 4$  and  $N^{\text{sic}} = 25$ . This indicates that the performance of the proposed SIC-based schemes becomes better when  $N^{\text{sic}}$  decreases. However, the selection of  $N^{\text{sic}}$  does not have an obvious impact on the detection performance when  $10 \leq N^{\text{sic}} \leq 20$ .

The comparison of the simulated results and theoretical SE is shown in Fig. 12. It can be observed that the simulated results can be well predicted by the SE results. Besides, the proposed two OAMP-MMV algorithms converge very fast, especially for the OAMP-MMV-ASL algorithm (in less than 10 iterations). The MSE gap between the OAMP-MMV-SSL algorithm and OAMP-MMV-ASL algorithm demonstrates the better detection performance of the latter algorithm.

## VI. CONCLUSIONS

In this paper, a slotted grant-free massive random access scheme was proposed for enabling mMTC in future massive IoT, where the JADD problem was investigated. Specifically, we assumed that the devices' activity remain unchanged in  $T$  successive OFDM symbols (i.e., one time slot), and the uplink massive access at the BS can be formulated as an MMV CS problem by utilizing the inherently structured sparsity. Different from the conventional OAMP algorithm limited to the SMV problem, the proposed OAMP-MMV-SSL and OAMP-MMV-ASL algorithms can capture the structured sparsity for improved JADD performance. Particularly, the former enjoys lower computational complexity, while the latter has a better activity detection performance, especially when  $T$  is small. Considering the practical application of our proposed algorithms, the OAMP-MMV-ASL algorithm can be applied when  $T$  is small, while the OAMP-MMV-SSL algorithm is better if  $T$  is large. To further improve the performance, SIC-based methods were integrated into the two proposed algorithms and channel coding was taken into account. In particular, the LLR obtained in soft-decision was applied to refine the estimation of the active devices. Besides, the SE was derived to predict the performance of our algorithms.

Simulation results confirmed the superiority of the proposed algorithms over state-of-the-art solutions, and demonstrated the gain of the proposed SIC-based JADD scheme, especially for the SIC-based OAMP-MMV-SSL algorithm.

## APPENDIX

### A. Derivation of EM Update

Firstly, we define an intermediate variable  $\mathbf{E} = \tilde{\mathbf{S}}\mathbf{X}$  for simplicity. Moreover,  $e_{m,t}$  denotes the  $(m,t)$ -th element of matrix  $\mathbf{E}$ , and  $\tilde{s}_{m,k}$  denotes the  $(m,k)$ -th element of matrix  $\tilde{\mathbf{S}}$ . In OAMP-MMV-SSL algorithm, we have the following relation

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Y}) &= \ln p(\mathbf{Y}|\mathbf{X}) + \ln p(\mathbf{X}) \\ &= \prod_{m=1}^M \prod_{t=1}^T \mathcal{CN}(y_{m,t}; e_{m,t}, \sigma^2) + \prod_{k=1}^K \prod_{t=1}^T p(x_{k,t}). \end{aligned} \quad (46)$$

The posterior distribution  $p(\mathbf{X}|\mathbf{Y}; \boldsymbol{\theta}^{i-1})$  can be approximated as

$$p(\mathbf{X}|\mathbf{Y}; \boldsymbol{\theta}^{i-1}) \approx \prod_{k=1}^K \prod_{t=1}^T p(x_{k,t}|r_{k,t}^i), \quad (47)$$

where  $p(x_{k,t}|r_{k,t}^i)$  is defined in (14).

Then, we derive the update of noise variance  $\sigma^2$  in OAMP-MMV-SSL algorithm. Substituting (46) into (19), we have

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{i-1}) &= -MT \ln \sigma^2 \\ &\quad + \sum_{m=1}^M \sum_{t=1}^T \frac{\mathbb{E}[|y_{m,t} - e_{m,t}|^2 | \mathbf{Y}; \boldsymbol{\theta}^{i-1}]}{\sigma^2}, \end{aligned} \quad (48)$$

where the irrelevant term with respect to  $\sigma^2$  has been ignored. By setting the derivative of (48) with respect to  $\sigma^2$  to zero, we obtain the update rule in (22) as

$$\begin{aligned} (\sigma^2)^i &= \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \mathbb{E}[|y_{m,t} - e_{m,t}|^2 | \mathbf{Y}; \boldsymbol{\theta}^{i-1}] \\ &= \frac{1}{MT} \sum_{m=1}^M \sum_{t=1}^T \left\{ |y_{m,t} - \mathbb{E}[e_{m,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}]|^2 \right. \\ &\quad \left. + \text{var}[e_{m,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}] \right\}, \end{aligned} \quad (49)$$

where  $p(\mathbf{X}|\mathbf{Y}; \boldsymbol{\theta}^{i-1})$  can be approximated by (47), and

$$\mathbb{E}[e_{m,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}] = \mathbb{E}\left[\sum_{k=1}^K \tilde{s}_{m,k} x_{k,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}\right] = \sum_{k=1}^K \tilde{s}_{m,k} \mu_{k,t}^i, \quad (50)$$

$$\text{var}[e_{m,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}] = \sum_{k=1}^K |\tilde{s}_{m,k}|^2 \text{var}[x_{k,t} | \mathbf{Y}; \boldsymbol{\theta}^{i-1}] = \frac{1}{K} \sum_{k=1}^K \gamma_{k,t}^i. \quad (51)$$

The calculation of (50)-(51) exploits the definition of the auxiliary matrix  $\mathbf{E}$  and partial DFT matrix  $\tilde{\mathbf{S}}$ .

The derivation of  $\lambda_{k,t}$  in (25) is described as follows. Since the Dirac function is discontinuous, we adopt the following

$$\frac{\partial \ln p(\mathbf{X}, \mathbf{Y})}{\partial \lambda_{k,t}} = \frac{\partial}{\partial \lambda_{k,t}} \left[ \sum_{k=1}^K \sum_{t=1}^T \ln p(x_{k,t}) \right] = \lim_{\epsilon \rightarrow 0^+} \underbrace{\frac{-\mathcal{CN}(x_{k,t}; 0, \epsilon) + (1/L) \sum_{l=1}^L \mathcal{CN}(x_{k,t}; a_l, \epsilon)}{(1 - \lambda_{k,t}) \mathcal{CN}(x_{k,t}; 0, \epsilon) + (\lambda_{k,t}/L) \sum_{l=1}^L \mathcal{CN}(x_{k,t}; a_l, \epsilon)}}}_{d(x_{k,t})}. \quad (53)$$

approximation:  $\mathcal{CN}(x; x_0, \epsilon) \rightarrow \delta(x - x_0)$ , when  $\epsilon \rightarrow 0^+$ . Under this assumption, by exchanging the order of the integral and the derivative, we have

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{i-1})}{\partial \lambda_{k,t}} = \mathbb{E} \left[ \frac{\partial \ln p(\mathbf{X}, \mathbf{Y})}{\partial \lambda_{k,t}} \middle| \mathbf{Y}; \boldsymbol{\theta}^{i-1} \right], \quad (54)$$

where the derivative can be simplified as equation (53). Based on (53), we obtain

$$\lim_{\epsilon \rightarrow 0^+} d(0) = \lim_{\epsilon \rightarrow 0^+} \frac{-\mathcal{CN}(0; 0, \epsilon)}{(1 - \lambda_{k,t}) \mathcal{CN}(0; 0, \epsilon)} = \frac{-1}{1 - \lambda_{k,t}}, \quad (55)$$

$$\lim_{\epsilon \rightarrow 0^+} d(a_l) = \lim_{\epsilon \rightarrow 0^+} \frac{\mathcal{CN}(a_l; a_l, \epsilon)}{\lambda_{k,t} \mathcal{CN}(a_l; a_l, \epsilon)} = \frac{1}{\lambda_{k,t}}, \quad \forall l. \quad (56)$$

Substituting (47) and (53)-(56) into (54), we have

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{i-1})}{\partial \lambda_{k,t}} = -\frac{1 - \pi_{k,t}^i}{1 - \lambda_{k,t}} + \frac{\pi_{k,t}^i}{\lambda_{k,t}}. \quad (57)$$

By setting (57) equal to zero, we obtain the update rule in (21).

Following similar steps, the update rule of OAMP-MMV-ASL algorithm can also be derived, which is omitted for brevity. However, the update of  $\lambda_k$ , given as follows

$$\lambda_k^i = \frac{\lambda_k^{i-1} \prod_{t=1}^T \eta_{k,t}^i}{(1 - \lambda_k^{i-1}) \prod_{t=1}^T (1 - \eta_{k,t}^i) + \lambda_k^{i-1} \prod_{t=1}^T \eta_{k,t}^i}, \quad (58)$$

heavily depends on the accuracy of each  $\eta_{k,t}$ , where  $\eta_{k,t}$  is given in (61). When some  $\eta_{k,t}$  is wrong due to the noise, the  $\lambda_k^i$  calculated by equation (58) can be opposite to the true activity pattern, leading to the deterioration of the detection performance. Therefore, we adopt the equation below

$$\begin{aligned} \lambda_k^i &= \frac{1}{T} \sum_{t=1}^T \xi_{k,t}^i \\ &= \frac{1}{T} \sum_{t=1}^T \frac{\lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}{(1 - \lambda_k^{i-1}) \prod_{t' \in \setminus t} (1 - \eta_{k,t'}^i) + \lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}, \end{aligned} \quad (59)$$

to mitigate the error caused by some inaccurate  $\eta_{k,t}$  and improve the robustness of the EM algorithm in our proposed OAMP-MMV-ASL algorithm.

### B. Derivation of Message Update for ASL Model

In the  $i$ -th iteration of OAMP-MMV-ASL algorithm, we start with the message through the path  $x_{k,t} \rightarrow f_{k,t} \rightarrow \alpha_k$ . Firstly, the message from variable node  $x_{k,t}$  to factor node

$f_{k,t}$  is  $m_{x_{k,t} \rightarrow f_{k,t}}^i = \mathcal{CN}(x_{k,t}; r_{k,t}^i, \tau_t^i)$ . Then, the message from factor node  $f_{k,t}$  to variable node  $\alpha_k$  can be written as

$$\begin{aligned} m_{f_{k,t} \rightarrow \alpha_k}^i &\propto \int m_{x_{k,t} \rightarrow f_{k,t}}^i f_{k,t}(x_{k,t}, \alpha_k) dx_{k,t} \\ &= \frac{1 - \alpha_k}{\pi \tau_t^i} \exp\left(-\frac{|r_{k,t}^i|^2}{\tau_t^i}\right) \\ &\quad + \frac{\alpha_k}{\pi \tau_t^i L} \sum_{l=1}^L \exp\left(-\frac{|r_{k,t}^i - a_l|^2}{\tau_t^i}\right) \\ &\propto (1 - \eta_{k,t}^i) \delta(\alpha_k) + \eta_{k,t}^i \delta(\alpha_k - 1), \end{aligned} \quad (60)$$

where  $\propto$  denotes equality up to a constant scale factor, and

$$\eta_{k,t}^i = 1 - \frac{1}{1 + \frac{1}{L} \sum_{l=1}^L q_{k,t,l}^i}. \quad (61)$$

The backward message is calculated through the path  $\bar{h}_k \rightarrow \alpha_k \rightarrow f_{k,t} \rightarrow x_{k,t}$ . The message from factor node  $\bar{h}_k$  to variable node  $\alpha_k$  is  $m_{\bar{h}_k \rightarrow \alpha_k}^i = (1 - \lambda_k^{i-1}) \delta(\alpha_k) + \lambda_k^{i-1} \delta(\alpha_k - 1)$ . Next, the message from variable node  $\alpha_k$  to factor node  $f_{k,t}$  is

$$\begin{aligned} m_{\alpha_k \rightarrow f_{k,t}}^i &\propto m_{\bar{h}_k \rightarrow \alpha_k}^i \prod_{t' \in \setminus t} m_{f_{k,t'}, \alpha_k}^i \\ &= (1 - \lambda_k^{i-1}) \prod_{t' \in \setminus t} (1 - \eta_{k,t'}^i) \delta(\alpha_k) \\ &\quad + \lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i \delta(\alpha_k - 1) \\ &\propto (1 - \xi_{k,t}^i) \delta(\alpha_k) + \xi_{k,t}^i \delta(\alpha_k - 1), \end{aligned} \quad (62)$$

where

$$\xi_{k,t}^i = \frac{\lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}{(1 - \lambda_k^{i-1}) \prod_{t' \in \setminus t} (1 - \eta_{k,t'}^i) + \lambda_k^{i-1} \prod_{t' \in \setminus t} \eta_{k,t'}^i}. \quad (63)$$

Finally, the message from factor node  $f_{k,t}$  to variable node  $x_{k,t}$  is

$$\begin{aligned} m_{f_{k,t} \rightarrow x_{k,t}}^i &\propto \sum_{\alpha_k} m_{\alpha_k \rightarrow f_{k,t}}^i f_{k,t}(x_{k,t}, \alpha_k) \\ &= (1 - \xi_{k,t}^i) \delta(x_{k,t}) + \frac{\xi_{k,t}^i}{L} \sum_{l=1}^L \delta(x_{k,t} - a_l). \end{aligned} \quad (64)$$

The approximate posterior distribution of  $x_{k,t}$  is given by

$$p(x_{k,t} | \mathbf{r}_k^i) \propto m_{f_{k,t} \rightarrow x_{k,t}}^i m_{x_{k,t} \rightarrow f_{k,t}}^i. \quad (65)$$

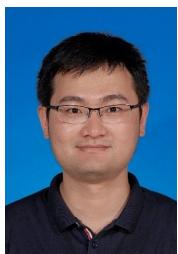
## REFERENCES

- [1] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sep. 2016.
- [2] F. Ghavimi and H. H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tut.*, vol. 17, no. 2, pp. 525-549, 2nd Quart. 2015.
- [3] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120-128, Feb. 2017.
- [4] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 1, pp. 4-16, 1st Quart. 2014.
- [5] 3GPP TS 38.300 V16.0.0, "Technical specification group radio access network; NR; NR and NG-RAN overall description," Dec. 2019.
- [6] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86-93, Jun. 2013.
- [7] J. Wang, Z. Zhang, and L. Hanzo, "Joint active user detection and channel estimation in massive access systems exploiting Reed-Muller sequences," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 739-752, Jun. 2019.
- [8] D. Zucchetto and A. Zanella, "Uncoordinated access schemes for the IoT: Approaches, regulations, and performance," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 48-54, Sept. 2017.
- [9] X. Chen, D. W. K. Ng, and W. Yu, "Massive access for 5G and beyond," [Online]. arXiv:2002.03491, 2020.
- [10] M. Shirvani-moghadam, M. Condoluci, M. Dohler, and S. J. Johnson, "On the fundamental limits of random non-orthogonal multiple access in cellular massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238-2252, Oct. 2017.
- [11] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74-81, Sept. 2015.
- [12] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454-465, Feb. 2011.
- [13] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972-974, Jul. 2012.
- [14] Y. Du *et al.*, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, Dec. 2017.
- [15] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320-2323, Nov. 2016.
- [16] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473-1476, Jul. 2016.
- [17] Y. Du *et al.*, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894-7909, Dec. 2018.
- [18] A. T. Abebe and C. G. Kang, "Iterative order recursive least square estimation for exploiting frame-wise sparsity in compressive sensing-based MTC," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1018-1021, May 2016.
- [19] X. Zhang, Y. Liang, and J. Fang, "Novel Bayesian inference algorithms for multiuser detection in M2M communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7833-7848, Sept. 2017.
- [20] B. K. Jeong, B. Shim, and K. B. Lee, "MAP-based active user and data detection for massive machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8481-8494, Sept. 2018.
- [21] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.
- [22] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 1993 Asilomar Conf.*, vol. 1, pp. 40-44.
- [23] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230-2249, May 2009.
- [24] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. Inf. Theory Workshop.*, Jan. 2010, pp. 1-5.
- [25] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211-244, 2001.
- [26] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 7th Conf. Uncertainty Artif. Intell.*, pp. 362-369, 2001.
- [27] J. Ma and L. Ping, "Orthogonal AMP," *IEEE Access*, vol. 5, pp. 2020-2033, 2017.
- [28] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658-4672, Oct. 2013.
- [29] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, Springer, 1998, pp. 355-368.
- [30] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498-519, Feb. 2001.
- [31] C. Bockelmann, "Iterative soft interference cancellation for sparse BPSK signals," *IEEE Commun. Lett.*, vol. 19, no. 5, pp. 855-858, May 2015.
- [32] J. Ahn, B. Shim, and K. B. Lee, "Sparsity-aware ordered successive interference cancellation for massive machine-type communications," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 134-137, Feb. 2018.
- [33] B. Knoop *et al.*, "Sparsity-aware successive interference cancellation with practical constraints," in *Proc. 17th Int. ITG Workshop Smart Antennas*, pp. 1-8, Mar. 2013.
- [34] J. M. Feng and C. H. Lee, "Generalized subspace pursuit for signal recovery from multiple-measurement vectors," in *Proc. Wireless Commun. Network Conf.*, pp. 2874-2878, Apr. 2013.
- [35] J. Rodrlguez-Fernndez, N. Gonzlez-Prelcic, K. Venugopal, and R. W. Heath, "Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2946-2960, May 2018.
- [36] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169-6183, Dec. 2015.
- [37] X. Lin, S. Wu, L. Kuang, Z. Ni, X. Meng, and C. Jiang, "Estimation of sparse massive MIMO-OFDM channels with approximately common support," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1179-1182, May 2017.
- [38] D. Hui, S. Sandberg, Y. Blankenship, M. Andersson, and L. Grosjean, "Channel coding in 5G new radio: A tutorial overview and performance comparison with 4G LTE," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 60-69, Dec. 2018.

**Yikun Mei** received the B.S. degree from the Beijing Institute of Technology, Beijing in 2019. He is currently pursuing the Master degree in the School of Information and Electronics of Beijing Institute of Technology, Beijing, China. His research interests include massive access for mMTC and sparse signal processing, etc.



**Zhen Gao** received the B.S. degree in information engineering from the Beijing Institute of Technology, Beijing, China, in 2011, and the Ph.D. degree in communication and signal processing with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China, in 2016. He is currently an Assistant Professor with the Beijing Institute of Technology. His research interests are in wireless communications, with a focus on multi-carrier modulations, multiple antenna systems, and sparse signal processing. He was a recipient of the IEEE Broadcast Technology Society 2016 Scott Helt Memorial Award (Best Paper), the Exemplary Reviewer of IEEE COMMUNICATION LETTERS in 2016, *IET Electronics Letters* Premium Award (Best Paper) 2016, and the Young Elite Scientists Sponsorship Program (2018-2021) from China Association for Science and Technology.



**Yongpeng Wu** (S'08–M'13–SM'17) received the B.S. degree in telecommunication engineering from Wuhan University, Wuhan, China, in July 2007, the Ph.D. degree in communication and signal processing with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, in November 2013.

Dr. Wu is currently a Tenure-Track Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China. Previously, he was senior research fellow with Institute for Communications Engineering, Technical University of Munich, Germany and the Humboldt research fellow and the senior research fellow with Institute for Digital Communications, University Erlangen-Nürnberg, Germany. During his doctoral studies, he conducted cooperative research at the Department of Electrical Engineering, Missouri University of Science and Technology, USA. His research interests include massive MIMO/MIMO systems, massive machine type communication, physical layer security, and signal processing for wireless communications.

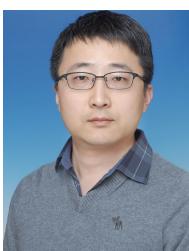
Dr. Wu was awarded the IEEE Student Travel Grants for IEEE International Conference on Communications (ICC) 2010, the Alexander von Humboldt Fellowship in 2014, the Travel Grants for IEEE Communication Theory Workshop 2016, the Excellent Doctoral Thesis Awards of China Communications Society 2016, the Exemplary Editor Award of IEEE Communication Letters 2017, and Young Elite Scientist Sponsorship Program by CAST 2017. He was an Exemplary Reviewer of the IEEE Transactions on Communications in 2015, 2016, 2018. He was the lead guest editor for the special issue "Physical Layer Security for 5G Wireless Networks" of the IEEE Journal on Selected Areas in Communications and lead guest editor of the special issue "Safeguarding 5G-and-Beyond Networks with Physical Layer Security" of IEEE Wireless Communications. He is currently an editor of the IEEE Wireless Communications, IEEE Transactions on Communications and IEEE Communications Letters. He has been a TPC member of various conferences, including Globecom, ICC, VTC, and PIMRC, etc.



**Derrick Wing Kwan Ng** (S'06–M'12–SM'17–F'21) received the bachelor degree with first-class honors and the Master of Philosophy (M.Phil.) degree in electronic engineering from the Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively. He received his Ph.D. degree from the University of British Columbia (UBC) in 2012. He was a senior postdoctoral fellow at the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany. He is now working as

a Senior Lecturer and a Scientia Fellow at the University of New South Wales, Sydney, Australia. His research interests include convex and non-convex optimization, physical layer security, IRS-assisted communication, UAV-assisted communication, wireless information and power transfer, and green (energy-efficient) wireless communications.

Dr. Ng received the Australian Research Council (ARC) Discovery Early Career Researcher Award 2017, the Best Paper Awards at the WCSP 2020, IEEE TCGCC Best Journal Paper Award 2018, INISCOM 2018, IEEE International Conference on Communications (ICC) 2018, 2021, IEEE International Conference on Computing, Networking and Communications (ICNC) 2016, IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, and the IEEE Third International Conference on Communications and Networking in China 2008. He has been serving as an editorial assistant to the Editor-in-Chief of the IEEE Transactions on Communications from Jan. 2012 to Dec. 2019. He is now serving as an editor for the IEEE Transactions on Communications, the IEEE Transactions on Wireless Communications, and an area editor for the IEEE Open Journal of the Communications Society. Also, he has been listed as a Highly Cited Researcher by Clarivate Analytics since 2018.



**Wei Chen** (M'13–SM'18) received the B.Eng. degree and M.Eng. degree from Beijing University of Posts and Telecommunications, China, in 2006 and 2009, respectively, and the Ph.D. degree in Computer Science from the University of Cambridge, UK, in 2013. Later, he was a Research Associate with the Computer Laboratory, University of Cambridge from 2013 to 2016. He is currently a Professor with Beijing Jiaotong University, Beijing, China. He is the recipient of the 2013 IET Wireless Sensor Systems Premium Award and the 2017 International Conference on Computer Vision (ICCV) Young Researcher Award. His current research interests include sparse representation, Bayesian inference, wireless communication systems and image processing.

Conference on Computer Vision (ICCV) Young Researcher Award. His current research interests include sparse representation, Bayesian inference, wireless communication systems and image processing.



**Marco Di Renzo** (Fellow, IEEE) received the Laurea (cum laude) and Ph.D. degrees in electrical engineering from the University of L'Aquila, Italy, in 2003 and 2007, respectively, and the Habilitation à Diriger des Recherches (Doctor of Science) degree from University Paris-Sud, France, in 2013. Since 2010, he has been with the French National Center for Scientific Research (CNRS), where he is a CNRS Research Director (CNRS Professor) with the Laboratory of Signals and Systems (L2S) of Paris-Saclay University – CNRS and CentraleSup-

elec, Paris, France. In Paris-Saclay University, he serves as the Coordinator of the Communications and Networks Research Area of the Laboratory of Excellence DigiCosme, and as a member of the Admission and Evaluation Committee of the Ph.D. School on Information and Communication Technologies. Currently, he serves as the Editor-in-Chief of IEEE Communications Letters, and is a Distinguished Speaker of the IEEE Vehicular Technology Society. In 2017–2020, he was a Distinguished Lecturer of the IEEE Vehicular Technology Society and IEEE Communications Society. Also, he served as an Editor and the Associate Editor-in-Chief of IEEE Communications Letters, and as an Editor of IEEE Transactions on Communications and IEEE Transactions on Wireless Communications. Currently, he serves as the Founding Chair of the Special Interest Group on Reconfigurable Intelligent Surfaces of the Wireless Technical Committee of the IEEE Communications Society, and is the Founding Lead Editor of the IEEE Communications Society Best Readings in Reconfigurable Intelligent Surfaces. He is a Highly Cited Researcher (Clarivate Analytics, 2019), a World's Top 2% Scientist from Stanford University (2020), a Fellow of the IEEE (2020), a Fellow of the IET (2020), an Ordinary Member of the EASA (2021), and a Member of the Academia Europaea (2021). He has received several individual distinctions and research awards, which include the IEEE Communications Society Best Young Researcher Award for Europe, Middle East and Africa, the Royal Academy of Engineering Distinguished Visiting Fellowship, the IEEE Jack Neubauer Memorial Best System Paper Award, the IEEE Communications Society Young Professional in Academia Award, the SEE-IEEE Alain Glavieux Award, the 2019 IEEE ICC Best Paper Award, the Nokia Foundation Visiting Professorship, and the 2021 EURASIP Journal on Wireless Communications and Networking Best Paper Award.



**Jun Zhang** received the B.S., M.S. and Ph.D. degrees in communications and electronic systems from Beihang University, Beijing, China, in 1987, 1991, and 2001, respectively. He used to be a professor in Beihang University, and has served as the dean of the school of Electronic and Information Engineering, the vice president, and the secretary of the Party Committee of Beihang University. Now he is a professor in Beijing Institute of Technology, and also the president of Beijing Institute of Technology.

His research interests are networked and collaborative air traffic management systems, covering signal processing, integrated and heterogeneous networks, and wireless communications. He has won the awards for science and technology in China many times, and he is a member of Chinese Academy of Engineering.